

Model Editing for Targeted Translation Error Repair

Danna Zheng^{1,2*}, Yang Liu¹, Zhu Liu¹, Vimal Bhat¹,
Manoj Aggarwal¹, Gerard Medioni¹, Najmeh Sadoughi¹

¹Amazon Prime Video,

²ILCC, School of Informatics, University of Edinburgh,

D.Zheng-6@sms.ed.ac.uk,

{yangnliu, zhuzliu, vimalb, manojagg, medioni, nnnourab}@amazon.com

Abstract

Machine translation systems still make systematic errors on idioms and other expressions that cannot be translated literally. Model editing offers a possible way to fix such recurring errors without retraining the whole model, but it has mostly been studied for monolingual factual updates rather than translation. We introduce MEMT-bench, the first benchmark for evaluating model editing in machine translation across multiple languages, models, and editing settings. The benchmark is built from idiomatic translation errors and tests whether an edit can correct a targeted mistranslation while preserving unrelated translation behavior. Experiments with causal tracing and representative editing methods reveal both the promise and difficulty of translation editing: edits can repair targeted errors, but maintaining cross-lingual locality and robustness under repeated edits remains challenging. Code and data are publicly available at [GitHub](#) and [Hugging Face](#), respectively.

1 Introduction

Large language models (LLMs) have advanced machine translation (MT) to a level comparable to junior or mid-level human translators (Yan et al., 2024). However, outputs often remain flawed due to overly literal translations (Yan et al., 2024). Existing approaches such as fine-tuning aim to improve translation quality globally (Cao et al., 2026; Hu et al., 2022; Khandelwal et al., 2021; Hu et al., 2019), but are not designed for precise corrections of specific recurring errors.

Model editing offers an alternative paradigm. Rather than retraining on large datasets, editing methods aim to adjust a model’s behavior with respect to a specific piece of knowledge while preserving its overall functionality (Mitchell et al.,

2022b). This makes editing appealing for targeted translation repair, where the goal is to fix systematic errors without affecting unrelated capabilities. Yet prior work has focused mainly on monolingual factual updates, such as modifying world knowledge in question answering (e.g., changing the British Prime Minister from Rishi Sunak to Keir Starmer). It remains unclear how editing behaves in cross-lingual settings, where a local change in one language pair may transfer, fail to transfer, or harm other translation directions.

Idiomatic expressions present a particularly demanding case: they resist literal translation, require context-sensitive understanding, and often lead to systematic mistranslations. These errors are especially useful for studying targeted repair because the desired correction is semantically specific but must still generalize across contexts and languages. Correcting such errors therefore requires fine-grained cross-lingual knowledge updates, making idioms a valuable test bed for examining model editing in translation.

To address this gap, we introduce MEMT-bench, the first benchmark for evaluating model editing in machine translation. Unlike a standard MT benchmark that ranks systems by absolute quality, MEMT-bench evaluates relative pre/post-edit behavior: whether an edit corrects a targeted mistranslation while preserving unrelated translation behavior. Each edit instance is accompanied by generalization and locality test sets that assess whether edits apply to new contexts while avoiding unintended side effects.

The paper is organized around three questions. First, we use causal tracing to identify where idiomatic translation knowledge is represented in QWEN2.5-3B-INSTRUCT. Second, we evaluate editing algorithms across reliability, generalization, and locality. Third, we test cross-lingual transfer and robustness under batch and sequential edits. Our results show a consistent repair-

*This work was done during an internship with Amazon Prime Video.

preservation trade-off: parameter-modifying methods repair targeted errors effectively but can damage locality, while parameter-preserving methods are more stable but less able to generalize edits.

2 Related Work

2.1 Machine Translation Error Repair

Recent advances in MT largely rely on fine-tuning with curated parallel data (Cao et al., 2026; Xu et al., 2024; Zhu et al., 2024a) or preferences (Jau-regi Unanue et al., 2026; Zhu et al., 2024b; He et al., 2024), achieving substantial benchmark gains. However, these methods operate at the system level and provide limited control over individual translation behaviors; correcting recurrent errors tied to specific constructions often affects unrelated inputs. An alternative line of work applies translation error correction (TEC) to post-edit MT outputs (Yang et al., 2025; Berger et al., 2024; Otake and Miyao, 2023). While effective at improving fluency and fixing localized errors, TEC leaves the underlying MT model unchanged, allowing the same errors to recur.

In contrast, model editing enables targeted and persistent parameter updates. We explore how model editing can be applied to MT to directly correct recurring translation errors without post-editing or retraining.

2.2 Model Editing

Model editing methods for LLMs can be broadly divided into parameter-modifying and parameter-preserving approaches. Parameter-modifying methods directly update model weights to encode new knowledge, including locate-then-edit techniques such as ROME (Meng et al., 2022), MEMIT (Meng et al., 2023), AlphaEdit (Fang et al., 2025), and EAMET (Dai et al., 2026) as well as meta-learning approaches that predict weight updates via hypernetworks (e.g., MEND (Mitchell et al., 2022a), MALMEN (Tan et al., 2024)). In contrast, parameter-preserving methods attach lightweight auxiliary modules to store new information without changing base parameters, enabling stable and repeated edits (e.g., T-Patcher (Huang et al., 2023), GRACE (Hartvigsen et al., 2023), WISE (Wang et al., 2024b)). Existing evaluations primarily focus on editing English factual knowledge expressed as subject–relation–object triples, including ZsRE (De Cao et al., 2021), CounterFact (Meng et al., 2022),

and MQuAKE (Zhong et al., 2023). Multilingual benchmarks such as CLICKER (Jiang et al., 2026), Bi-ZsRE (Wang et al., 2024a) and MLaKE (Wei et al., 2025) are derived by translating English triples, while recent work on unstructured editing (Jiang et al., 2025; Deng et al., 2025) remains largely English-centric.

In contrast, we study model editing for machine translation, where the edited knowledge is inherently cross-lingual rather than mono-lingual facts. To our knowledge, this is the first systematic investigation of model editing in MT.

3 Problem Formulation

Let $\mathcal{L} = \{\ell_0, \ell_1, \dots, \ell_{|T|}\}$ be a set of languages, where ℓ_0 is the source language and $\ell_t \in \{\ell_1, \dots, \ell_{|T|}\}$ is a target language. We consider a translation model $\mathcal{F}_\theta$. Given a source sentence x^{ℓ_0} and target language ℓ_t , the model generates $\mathcal{F}_\theta(x^{\ell_0}, \ell_t)$. An **editing descriptor** is $(x_e^{\ell_0}, \ell_t, y_e^{\ell_t})$ where the model fails to produce the desired output: $\mathcal{F}_\theta(x_e^{\ell_0}, \ell_t) \neq y_e^{\ell_t}$. An **editing algorithm** $\mathcal{E}$ takes $\mathcal{F}_\theta$ and descriptors $\mathcal{S}$, and returns an edited model $\mathcal{F}'_\theta = \mathcal{E}(\mathcal{F}_\theta, \mathcal{S})$. In the single-edit case $\mathcal{S} = \{(x_e^{\ell_0}, \ell_t, y_e^{\ell_t})\}$, the goal is $\mathcal{F}'_\theta(x_e^{\ell_0}, \ell_t) = y_e^{\ell_t}$.

A successful edit should localize its effect: it must change the model’s behavior only on inputs within the editing scope while leaving all unrelated predictions intact. Formally,

$$\mathcal{F}'_\theta(x) = \begin{cases} y_e, & x \in \mathcal{I}(x_e^{\ell_0}, y_e^{\ell_t}) \\ \mathcal{F}_\theta(x), & x \in \mathcal{O}(x_e^{\ell_0}, y_e^{\ell_t}) \end{cases} \quad (1)$$

where $\mathcal{I}(x_e^{\ell_0}, y_e^{\ell_t})$ denotes the set of in-scope examples that share the same error pattern, $\mathcal{O}(x_e^{\ell_0}, y_e^{\ell_t})$ denotes out-of-scope examples that are unrelated to the edit.

To characterize the quality of an edit, we evaluate the post-edited model $\mathcal{F}'_\theta$ along three key dimensions: **reliability**, **generalization**, and **locality**. Each assess both within the edited language pair ($\ell_0 \rightarrow \ell_t$, i.e., in-pair) and across other target languages ($\ell_0 \rightarrow \ell_{t'}, t' \neq t$, i.e., out-of-pair).

- **Reliability:** The extent to which the edit succeeds on the specific correction itself (in-pair) and whether the improvement transfers when translating the same source sentence into other languages (out-of-pair).
- **Generalization:** The ability of the edit to correctly extend to in-scope examples who share

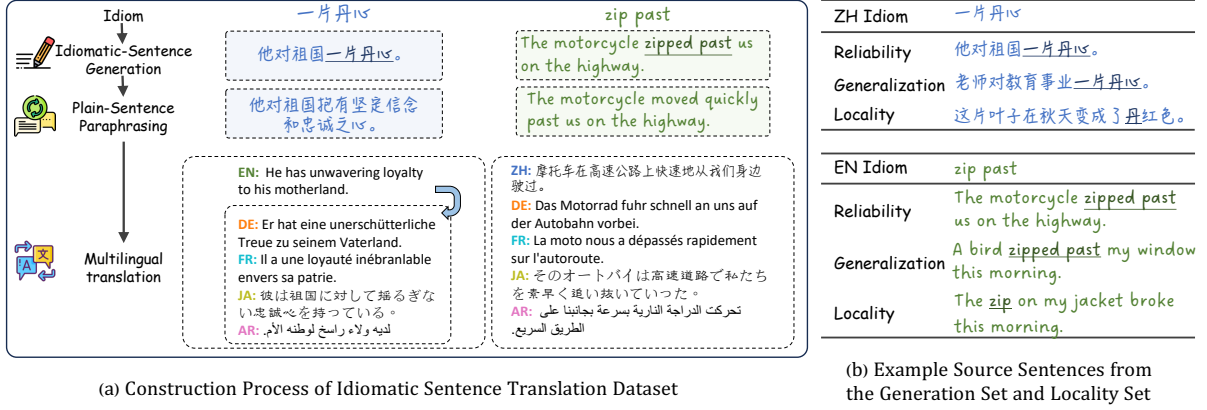


Figure 1: The figure illustrates the construction process of IdiomMT in MEMT-bench.

the same error pattern as the editing descriptor while occurring in different contexts.

- **Locality:** The degree to which the edit preserves model performance on out-of-scope examples.

4 MEMT-bench

We introduce MEMT-bench, a multilingual benchmark for model editing in translation. Unlike a conventional MT benchmark, MEMT-bench evaluates relative pre/post-edit behavior: whether an edit moves an erroneous output toward a controlled target while preserving unrelated translations. Additional design clarifications are in Appendix B. It is built from realistic correction scenarios through three stages: 1) Parallel idiom translation dataset creation, 2) MT evaluation on idiom translation, 3) Error filtering for editing the model.

4.1 Parallel Idiom Translation Dataset Creation

To address the scarcity of high-quality idiom translation data, we construct two large-scale parallel datasets: IdiomMT-en2x and IdiomMT-zh2x, covering translations from English and Chinese into multiple target languages (English, Chinese, Japanese, Arabic, French, and German).

4.1.1 Construction Process

Figure 1 illustrates the creation pipeline, which uses a plain-English pivot to ensure alignment and reduce noise from rare idiomatic structures. Both datasets are generated using LLMs based on IdiomKB (Li et al., 2024), which contains 3,990 English and 8,643 Chinese idioms with definitions.

Step 1: Idiom sentence generation For each idiom u , an LLM generates a short, natural sentence s_u^i in which u is explicitly used and constitutes the primary semantic challenge.

Step 2: Plain paraphrasing Each idiom sentence is paraphrased into a non-idiomatic plain sentence s_u^p that preserves the original meaning. The plain sentence is used only as a meaning-preserving pivot for reference construction, not as the final evaluation target.

Step 3: Multilingual translation Plain sentences are translated into five target languages. English sentences are translated directly; Chinese sentences are first translated into English and then into the remaining languages, using English as a pivot to reduce semantic drift and improve consistency, especially for low-resource language pairs.

Using the same pipeline with alternative sentence-generation prompts, we also construct two auxiliary test sets: a *generalization* set with idioms in varied contexts to test in-scope generalization, and a *locality* set with semantically unrelated sentences with partial lexical overlap to detect unintended side effects.

All data are generated using CLAUDE-SONNET-4. Prompt are in Table 5, 6, 7 in Appendix A.

4.1.2 Quality Control and Data Filtering

We apply three automatic quality-control steps:

Idiom Usage Verification We first locate idiom spans via string matching. For cases involving internal modification or discontinuity, we apply an extraction prompt (Table 8). Samples failing both checks are discarded.

Source: English				
Metric	m2m100-1.2B	nllb200-3.3B	Qwen2.5-3B	Qwen2.5-7B
COMET (↑)	0.691	0.778	0.744	0.772
BLEURT-20 (↑)	0.496	0.584	0.542	0.586
MetricX-23 (↓)	2.869	1.796	1.853	1.773
Claude-Scale (↑)	43.207	63.536	55.781	64.312
Claude-Binary (↑)	0.165	0.408	0.311	0.430

Source: Chinese				
Metric	m2m100-1.2B	nllb200-3.3B	Qwen2.5-3B	Qwen2.5-7B
COMET (↑)	0.710	0.760	0.756	0.779
BLEURT-20 (↑)	0.555	0.609	0.570	0.621
MetricX-23 (↓)	2.009	1.343	1.353	1.481
Claude-Scale (↑)	40.078	52.155	57.149	64.380
Claude-Binary (↑)	0.117	0.248	0.346	0.441

Table 1: Average translation performance on IdiomMT-en2x/zh2x (darker = larger within row)

Paraphrase Prompt Validation To assess whether our prompt reliably produces meaning-preserving paraphrases, we evaluate it on the PIE dataset (Zhou et al., 2021), which contains gold paraphrases of idiomatic sentences. Our approach achieves a BLEU score of 0.54, indicating strong semantic fidelity and fluency.

Translation Quality Estimation We use **MetricX-23-QE-XL**, a strong reference-free MT quality estimator (Perrella et al., 2024), to score each plain sentence–translation pair on a 0–25 scale (lower is better). We retain only samples with scores below 1.5, well under the practical quality threshold of 3.57 (Perrella et al., 2024).

Besides, idioms lacking three validated generalization or locality samples are removed to maintain consistency and balance.

4.1.3 Data Statistic

Each instance in the final dataset includes: the source idiom u (English or Chinese), one idiom-containing sentence s_u^i , reference translations of s_u^i in five target languages, three generalization samples with references in five target languages, and three locality samples with references in five target languages.

IdiomMT-en2x contains 3,133 English-to-X instances, and IdiomMT-zh2x contains 6,447 Chinese-to-X instances. Example entries are provided in Table 22 and 23 in Appendix G.

4.2 Collection of MT outputs and Evaluation

We collect translations from four popular MT systems across two idiom datasets.

Models: Two widely used multilingual MT systems (M2M100-1.2B, NLLB-200-3.3B) and two general-purpose LLMs (QWEN2.5-INSTRUCT-3B,

Source: English				
Target	m2m100-1.2B	nllb200-3.3B	Qwen2.5-3B	Qwen2.5-7B
Chinese	2935	2495	2061	1890
Japanese	2980	2605	2653	2493
German	2971	2416	2674	2558
French	2969	2417	2695	2436
Arabic	3081	2643	2852	2547
All Targets	2738	1810	1594	1328

Source: Chinese				
Target	m2m100-1.2B	nllb200-3.3B	Qwen2.5-3B	Qwen2.5-7B
English	5284	4175	2681	2428
Japanese	6008	5805	5558	5511
German	5819	5182	4891	4209
French	5712	5016	4709	3887
Arabic	6285	5421	5958	5201
All Targets	4807	3518	1928	1514

Table 2: Editing Samples Statistic

QWEN2.5-INSTRUCT-7B). They provide diverse error descriptors rather than a ranking of state-of-the-art MT systems; the benchmark can be applied to additional translators.

Settings: All models use beam size 5 and 200 maximum new tokens. Prompts for LLMs are in Table 7 in Appendix A.

Metrics: We evaluate with three complementary automatic metrics: **MetricX-23-XL** (0–25, lower is better), **BLEURT-20** (0–1+, higher is better), and **COMET** (0–1, higher is better). Additionally, we use two LLM-based judgments (Huang et al., 2024): **Claude-Scale** (0–100, higher is better) and **Claude-BI** (0 or 1, higher is better), both based on **CLAUDE-SONNET-4**. Prompt templates are listed in Table 9 in Appendix A.

Translators Performance: As shown in Table 1, QWEN2.5-7B leads overall, with 3B trailing slightly behind NLLB. All models struggle with idioms (Claude-Scale < 65, Claude-Bi < 0.45), confirming this as a core challenge for editing. Appendix C gives detailed results.

4.3 Filtering Erroneous Translations for Editing

To construct reliable editing descriptors, we filter out low-quality translations using a combination of metric-based quality thresholds and contrastive evaluation. A translation is considered erroneous if it is flagged by either of these criteria.

Quality Metric Thresholds A translation is flagged if any of the following holds: COMET < 0.7, BLEURT-20 < 0.6, MetricX-23-XL > 4, ClaudeScale < 50, or ClaudeBI = 0. These strict thresholds, following the empirical practice of (Per-

rella et al., 2024), effectively detect severe semantic and fluency errors.

Contrastive Quality Checks To target idiom errors, we generate literal (word-by-word) translations using CLAUDE-SONNET-4 with prompt in Table 10 (Appendix A) and compare system outputs against both reference and literal using COMET, BLEURT, MetricX, and ClaudeScale. Outputs closer to the literal are flagged.

Statistic of Editing Descriptors Table 2 reports filtered samples per model and language pair. Each model yields over 1,000 consistent errors across all target languages. Examples are in Tables 24 and 25 in Appendix G.

5 Model Editing Exploration

We focus on QWEN2.5-INSTRUCT-3B as a representative model to enable controlled and in-depth analysis of translation-specific model editing. We address three research questions:

RQ1: How does translation knowledge propagate across layers?

RQ2: How effectively do existing editing methods repair translation errors?

RQ3: How robust are these methods under multiple translation edits?

5.1 RQ1: How does translation knowledge propagate across layers?

We investigate where idiom translation knowledge is encoded across layers, aiming to identify effective intervention points for model editing.

5.1.1 Method

Following Meng et al. (2022), we estimate the causal importance of internal states via three forward passes:

Clean Run: Given a source sentence x and its correct model-generated translation y , we extract hidden representations $h_i^{(l)}$ for each token i at each layer l . We then compute the average log probability of the translation: $\text{avg log } P[y] = \frac{1}{|y|} \sum_{j=1}^{|y|} \log P(y_j | x)$. Unlike prior work (Meng et al., 2022), which focuses only on the first token, we average across the full sequence, as idiomatic meaning is distributed rather than front-loaded.

Corrupted Run: To disrupt idiom understanding, we inject Gaussian noise into the layer-0 embeddings of idiom tokens: $h_i^{(0)} \leftarrow h_i^{(0)} + \epsilon$, $\epsilon \sim$

$\mathcal{N}(0, \nu)$, where ν is set to three times the empirical standard deviation of the embedding space, following Meng et al. (2022). The resulting average log probability is $\text{avg log } P^*[y]$.

Corrupted-with-Restoration Run: For a given state $h_i^{(l)}$, we restore its clean value during the corrupted run and recompute the average log probability, denoted $\text{avg log } P_{\text{clean } h_i^{(l)}}^*[y]$.

The causal contribution of each state is measured by its *Indirect Effect* (IE):

$$\text{IE}(h_i^{(l)}) = \text{avg log } P_{\text{clean } h_i^{(l)}}^*[y] - \text{avg log } P^*[y] \quad (2)$$

We report Average Indirect Effect (AIE) by averaging IE over samples.

5.1.2 Data Selection

We select translations where QWEN2.5-INSTRUCT-3B achieves near-human performance (COMET ≥ 0.7 , BLEURT-20 ≥ 0.8 , MetricX-23-XL ≤ 1.5 , ClaudeScale ≥ 80 , ClaudeBI = 1), and that are closer to references than to literal negatives. Up to 100 idiomatic samples per language pair are used.

5.1.3 Results

Figure 2 shows layer-wise IE patterns for English $\rightarrow$ X and Chinese $\rightarrow$ X. Using the residual decomposition $h_i^{(l)} = h_i^{(l-1)} + a_i^{(l)} + m_i^{(l)}$, we isolate attention and MLP contributions.

MLP-Dominated Pathways Across both directions, attention contributions concentrate in early layers and decay quickly, while MLP effects peak in early-to-middle layers and closely track the hidden-state IE. This indicates that MLP activations serve as the most reliable and controllable carriers of idiomatic semantics, suggesting early mid-layer MLPs as stable intervention points (around layer~7 for English $\rightarrow$ X and layer~5 for Chinese $\rightarrow$ X).

Token-Level Effects At the token level, the final idiom token consistently exhibits the largest indirect effect, indicating the point at which idiomatic meaning is consolidated prior to generation.

Cross-Lingual Depth Asymmetry. The depth at which MLP influence is maximal varies by source language. English $\rightarrow$ X shows deeper and more distributed MLP peaks (layers 7–15), whereas Chinese $\rightarrow$ X peaks earlier and decays faster. We treat this as an empirical pattern rather

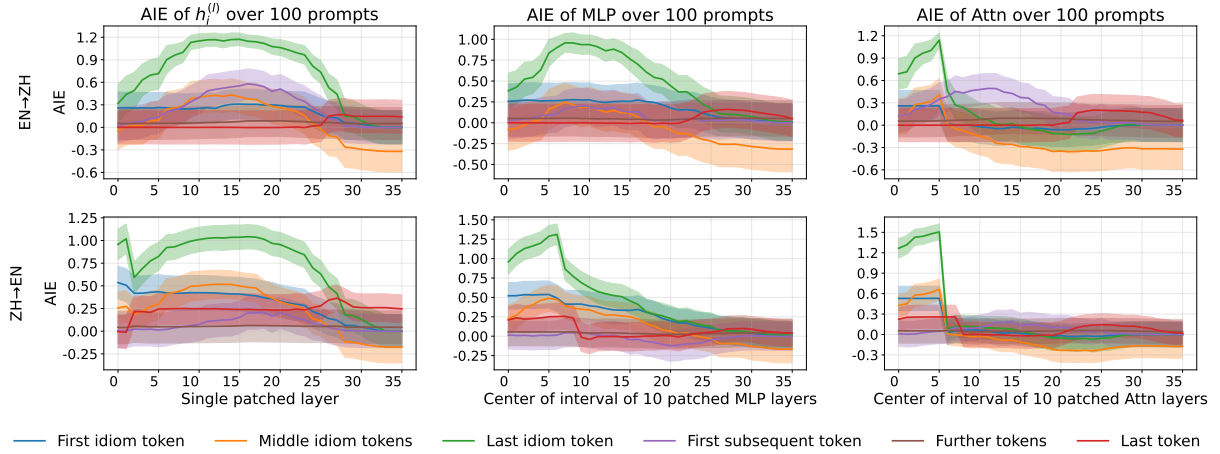


Figure 2: Mean causal traces of Qwen2.5-Instruct-3B over 100 translation samples for English→Chinese and Chinese→English, with 95% confidence intervals. Similar patterns are observed for other translation pairs sourced from English or Chinese. Full visualizations, including heatmaps and line plots, are provided in Appendix D.

than a causal linguistic claim; tokenization, morphology, and idiom lexicalization may all contribute.

5.2 RQ2: How effectively do existing editing methods repair translation errors?

We explore applying factual model editing to translation tasks. Focusing on a single-edit setting, each edit corrects one specific idiomatic mistranslation. This controlled setup isolates the effect of individual edits and helps clarify their intrinsic behavior before extending to multi-edit scenarios in RQ3.

5.2.1 Experimental Setup

We evaluate single edits on two directions: English→Chinese and Chinese→English.

Editing Methods We evaluate parameter-modifying methods including FT-M (Wang et al., 2024c), ROME (Meng et al., 2022), MEMIT (Meng et al., 2023), AlphaEdit (Fang et al., 2025), and UNKE (Deng et al., 2025); and parameter-preserving methods including GRACE (Hartvigsen et al., 2023) and WISE (Wang et al., 2024b). Method details are provided in Appendix E.

Data Selection We select 1,000 editing instances from Section 4.3 that fail across all target languages. To assess locality more broadly, we additionally include 1,000 FLORES translation samples per direction (English→Chinese, German, French, Japanese, Arabic and Chinese→English, German, French, Japanese, Arabic) to test whether edits affect general translation quality. To evaluate interference with broader reasoning and factual

abilities, we further sample 1,000 MMLU samples across six languages (Chinese, English, German, French, Japanese, and Arabic).

Hyperparameters For parameter-modifying methods, we edit layer 7 for English-source and layer 5 for Chinese-source edits, following Section 5.1. For multi-layer editors (MEMIT, AlphaEdit), we update layers 5–9 for English and 2–6 for Chinese, covering the core idiomatic region and observed depth asymmetry. For parameter-preserving methods, auxiliary modules are placed at layer 30 of Qwen2.5-3B-Instruct, following Wang et al. (2024b). Other hyperparameters follow EasyEdit (Wang et al., 2024c).

Metrics We report average scores across evaluation sets, with gain rates for reliability and generalization and preservation rates for locality. For translation tasks, we apply **FastText-based language detection**: outputs in the correct target language are scored with BLEURT, while mismatched outputs receive 0. We use BLEURT for efficient paired pre/post-edit comparison; Section 4.3 uses multiple metrics to construct high-confidence descriptors. For MMLU, we follow the **official protocol**, selecting the option with the highest predicted probability and assigning 1 if correct and 0 otherwise. Gain rate is the fraction of samples improved by editing, and preservation rate is the fraction not degraded.

5.2.2 Results

Parameter-modifying methods enable strong repair As shown in Table 3, parameter-modifying methods achieve near-ceiling in-pair reliability and

	Reliability				Generalization				Locality				Flores				MMLU	
	In-Pair		Out-of-Pair		In-Pair		Out-of-Pair		In-Pair		Out-of-Pair		In-Pair		Out-of-Pair		Average	
	Score	Gain	Score	Gain	Score	Gain	Score	Gain	Score	Pres.	Score	Pres.	Score	Pres.	Score	Pres.	Score	Pres.
<i>Editing Descriptor: English to Chinese</i>																		
pre-edit	0.511	-	0.434	-	0.542	-	0.457	-	0.767	-	0.666	-	0.694	-	0.615	-	0.548	-
FT	0.954	0.987	0.502	0.619	0.651	0.706	0.522	0.607	0.749	0.527	0.647	0.571	0.682	0.496	0.613	0.553	0.546	0.953
ROME	0.961	0.999	0.465	0.536	0.625	0.654	0.472	0.497	0.756	0.769	0.655	0.793	0.694	0.857	0.614	0.825	0.549	0.995
MEMIT	0.938	0.987	0.487	0.581	0.640	0.690	0.485	0.528	0.756	0.766	0.658	0.802	0.694	0.862	0.615	0.819	0.549	0.991
AlphaEdit	0.963	0.998	0.474	0.561	0.641	0.679	0.475	0.510	0.752	0.741	0.653	0.784	0.694	0.850	0.616	0.825	0.550	0.994
UNKE	0.942	0.982	0.365	0.382	0.557	0.320	0.437	0.264	0.750	0.746	0.636	0.739	0.692	0.843	0.614	0.785	0.549	0.988
GRACE	0.917	0.961	0.434	0.083	0.541	0.060	0.457	0.089	0.768	0.905	0.666	0.932	0.694	0.888	0.615	0.856	0.548	1.000
WISE	0.840	0.816	0.424	0.268	0.606	0.509	0.450	0.226	0.754	0.731	0.654	0.803	0.689	0.775	0.608	0.763	0.548	0.990
<i>Editing Descriptor: Chinese to English</i>																		
pre-edit	0.593	-	0.436	-	0.645	-	0.465	-	0.801	-	0.626	-	0.744	-	0.560	-	0.548	-
FT	0.971	0.983	0.460	0.548	0.722	0.743	0.465	0.506	0.763	0.552	0.560	0.495	0.699	0.494	0.502	0.472	0.538	0.960
ROME	0.980	0.997	0.415	0.447	0.701	0.660	0.424	0.406	0.796	0.775	0.619	0.767	0.746	0.850	0.558	0.761	0.547	0.993
MEMIT	0.885	0.899	0.384	0.417	0.658	0.587	0.391	0.371	0.754	0.659	0.561	0.642	0.705	0.641	0.507	0.580	0.508	0.918
AlphaEdit	0.936	0.970	0.420	0.456	0.689	0.624	0.422	0.408	0.781	0.631	0.605	0.618	0.742	0.690	0.554	0.611	0.534	0.961
UNKE	0.969	0.985	0.332	0.330	0.655	0.324	0.437	0.282	0.794	0.774	0.611	0.763	0.744	0.858	0.561	0.759	0.543	0.985
GRACE	0.966	0.982	0.436	0.094	0.645	0.079	0.465	0.100	0.800	0.860	0.625	0.891	0.743	0.930	0.560	0.836	0.548	1.000
WISE	0.805	0.757	0.405	0.282	0.674	0.470	0.430	0.236	0.781	0.717	0.583	0.701	0.727	0.716	0.523	0.640	0.544	0.983

Table 3: Performance of different editing methods on translation error repair. Pre-edit denotes the original Qwen2.5-3B-Instruct model. Gain and Pres indicate gain and preserve rates, where darker green denotes higher than 0.5 and darker red denotes lower than 0.5. For scores, green (red) marks improvement (degradation) relative to pre-edit, with darker shades indicating larger magnitude. Color intensity is normalized per column.

strong in-pair generalization, substantially outperforming the pre-edit baseline. Among them, FT also performs well on out-of-pair repair, indicating effective cross-lingual transfer. Locate-then-edit methods such as MEMIT and AlphaEdit achieve repair performance comparable to FT, particularly for English-source edits.

Parameter-preserving methods maintain global stability Among parameter-preserving methods, GRACE achieves the highest preservation of general translation quality and factual reasoning, owing to its adaptor-based isolation of edits. In contrast, WISE, while safer than FT, shows weaker preservation and limited gains, likely due to instability in its trainable routing mechanism.

Directional asymmetry Edits transfer less effectively in Chinese→English than in English→Chinese translation, particularly in out-of-pair reliability and generalization. This asymmetry reflects language-specific differences in semantic compression: Chinese idioms encode denser semantic content, making Chinese→English editing more challenging.

The repair–preservation trade-off defines practical performance A clear trade-off emerges: FT yields the strongest repair gains but suffers the largest locality degradation, whereas GRACE offers maximal stability with limited repair capacity. Locate-then-edit methods such like ROME and AlphaEdit strike a more balanced compromise,

though precise and efficient translation editing remains an open challenge.

5.3 RQ3: How robust are these methods under multiple translation edits?

We evaluate robustness under batch and sequential translation edits. These are editing-time (not inference-time) settings: batch editing updates the model once with multiple descriptors, while sequential editing applies descriptors one by one to the edited model.

5.3.1 Experimental Setup

We evaluate sequential editing after 1, 4, 16, 64, 256, and 1024 edits (one edit per step), and batch editing with batch sizes of 1, 4, 16, 32, 64, 128, and 256. All experiments are conducted on a single A100-SXM4-40GB GPU. Due to memory constraints, FT supports batch sizes up to 2^5 , while UNKE supports up to 2^7 .

Editing Methods All methods are evaluated under sequential editing. Batch editing is tested for FT, MEMIT, AlphaEdit, and UNKE, as ROME, GRACE, and WISE lack batch-update support.

Data Used We sample eight subsets of editing descriptors from translation instances that fail across all target languages (Section 4.3) and report results averaged over the subsets.

Hyperparameters and Metrics Hyperparameters and evaluation metrics follow the single-edit

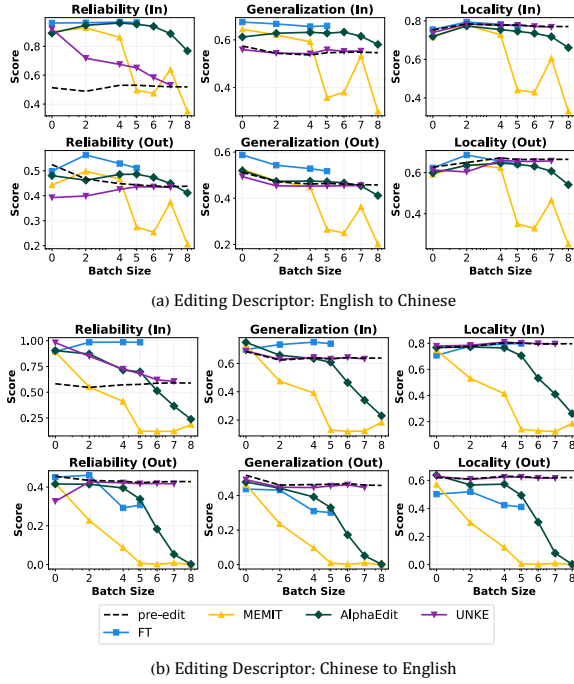


Figure 3: Editing performance versus batch size for English-to-Chinese and Chinese-to-English descriptor editing. The x-axis shows batch sizes as powers of two. “In” refers to in-pair language direction. “Out” refers to out-of-pair language direction.

setting described in Section 5.2.

5.3.2 Results

Batch Editing Performance Figures 3 show that performance degrades as batch size increases across all methods, with varying severity. FT performs best on small batches but degrades beyond 2^2 , indicating edit interference. AlphaEdit shows milder degradation with lower overall reliability. UNKE remains stable on locality and generalization but suffers declining in-pair reliability, while MEMIT degrades most sharply across all metrics, likely due to key-value conflicts.

Sequential Editing Performance As shown in Figure 4, parameter-modifying methods degrade cumulatively as edits accumulate. FT and UNKE drop fastest, with sharp out-of-pair reliability losses after a few steps; ROME and MEMIT decline more gradually, and AlphaEdit is the most stable among them. In contrast, parameter-preserving methods are more robust: GRACE maintains high in-pair reliability and stable overall performance with minor degradation, indicating effective mitigation of cumulative interference.

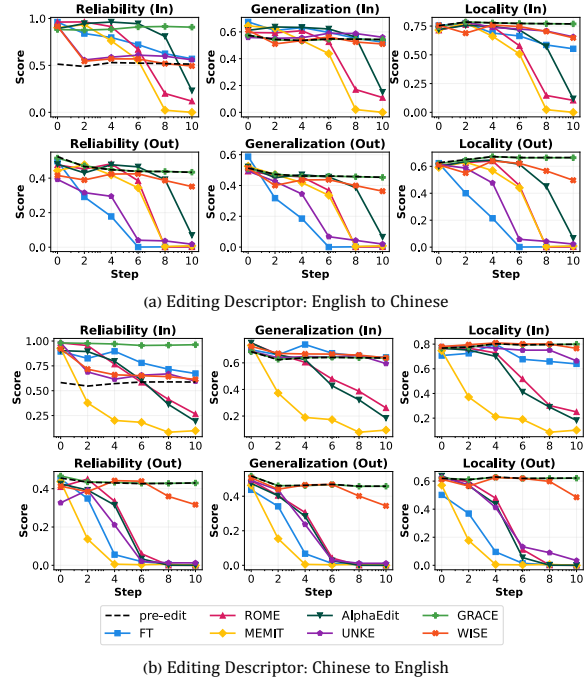


Figure 4: Editing performance versus sequential step for English-to-Chinese and Chinese-to-English descriptor editing. The x-axis represents steps in powers of two.

Directional Asymmetry Consistent with single-editing results, Chinese→English editing is less stable than English→Chinese, degrading earlier in both batch and sequential settings.

5.4 Case Analysis

Representative successful and failed cases are presented in Appendix H. We observe that parameter-modifying methods frequently exhibit side effects such as instruction repetition and nonsensical outputs. Unintended language shifts occur particularly under large batch sizes and sequential editing, even with small update steps (Figure 13). It remains unclear whether these issues arise from multilingual forgetting or degraded instruction-following. By contrast, parameter-preserving methods such as GRACE show limited generalization but largely maintain original predictions, resulting in minimal gains with minimal harm.

6 Conclusion

This work introduces MENT-bench, the first benchmark for model editing in machine translation. Experiments show that idiomatic translation knowledge concentrates in early mid-layer MLPs, enabling effective intervention. Parameter-modifying methods provide stronger repairs, while

parameter-preserving methods better maintain global stability. Future work should improve locality and stability for parameter-modifying methods and enhance generalization and batch-editing support for parameter-preserving approaches.

Limitations

This work focuses on idiomatic translation errors as a test bed for studying model editing in machine translation. While idioms represent a linguistically challenging and semantically rich error type, they do not cover other common MT errors such as syntactic reordering, discourse-level inconsistencies, or domain-specific terminology, and our findings may not fully generalize to these cases.

Our editing experiments are conducted primarily on Qwen2.5-Instruct-3B due to computational constraints and the high cost of large-scale batch and sequential editing. Although MEMT-bench itself is model-agnostic and includes outputs from multiple MT systems, editing behavior may differ across architectures, scales, and newer specialized translation LLMs.

Ethical Considerations

This work does not involve human subjects or personal data. All data used in MEMT-bench are derived from publicly available idiom resources and generated by large language models.

The model editing techniques studied in this paper are intended to correct systematic translation errors and improve model reliability. However, inappropriate or uncontrolled edits may introduce unintended side effects, such as performance degradation in unrelated languages or tasks. By explicitly evaluating locality, generalization, and stability under batch and sequential editing settings, this work aims to make both the benefits and the risks of multilingual model editing more transparent.

References

- Nathaniel Berger, Stefan Riezler, Miriam Exel, and Matthias Huck. 2024. [Prompting large language models with human error markings for self-correcting machine translation](#). In *Proceedings of the 25th Annual Conference of the European Association for Machine Translation (Volume 1)*, pages 636–646, Sheffield, UK. European Association for Machine Translation (EAMT).
- Zhe Cao, Yusuke Oda, Qianying Liu, Akiko Aizawa, and Taro Watanabe. 2026. [Completely modular fine-tuning for dynamic language adaptation](#). In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 4828–4845, Rabat, Morocco. Association for Computational Linguistics.
- Yanbo Dai, Zhenlan Ji, Zongjie Li, and Shuai Wang. 2026. [EAMET: ROBUST MASSIVE MODEL EDITING VIA EMBEDDING ALIGNMENT OPTIMIZATION](#). In *The Fourteenth International Conference on Learning Representations*.
- Nicola De Cao, Wilker Aziz, and Ivan Titov. 2021. [Editing factual knowledge in language models](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6491–6506, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Jingcheng Deng, Zihao Wei, Liang Pang, Hanxing Ding, Huawei Shen, and Xueqi Cheng. 2025. [Everything is editable: Extend knowledge editing to unstructured data in large language models](#). In *The Thirteenth International Conference on Learning Representations*.
- Junfeng Fang, Houcheng Jiang, Kun Wang, Yunshan Ma, Jie Shi, Xiang Wang, Xiangnan He, and Tat-Seng Chua. 2025. [Alphaedit: Null-space constrained knowledge editing for language models](#). In *The Thirteenth International Conference on Learning Representations*.
- Thomas Hartvigsen, Swami Sankaranarayanan, Hamid Palangi, Yoon Kim, and Marzyeh Ghassemi. 2023. [Aging with GRACE: Lifelong model editing with discrete key-value adapters](#). In *Thirty-seventh Conference on Neural Information Processing Systems*.
- Zhiwei He, Xing Wang, Wenxiang Jiao, Zhuosheng Zhang, Rui Wang, Shuming Shi, and Zhaopeng Tu. 2024. [Improving machine translation with human feedback: An exploration of quality estimation as a reward model](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8164–8180, Mexico City, Mexico. Association for Computational Linguistics.
- Junjie Hu, Hiroaki Hayashi, Kyunghyun Cho, and Graham Neubig. 2022. [DEEP: DENOISING ENTITY PRE-TRAINING FOR NEURAL MACHINE TRANSLATION](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1753–1766, Dublin, Ireland. Association for Computational Linguistics.
- Junjie Hu, Mengzhou Xia, Graham Neubig, and Jaime Carbonell. 2019. [Domain adaptation of neural machine translation by lexicon induction](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2989–3001, Florence, Italy. Association for Computational Linguistics.
- Xu Huang, Zhirui Zhang, Xiang Geng, Yichao Du, Jijun Chen, and Shujian Huang. 2024. [Lost in the](#)

- source language: How large language models evaluate the quality of machine translation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3546–3562, Bangkok, Thailand. Association for Computational Linguistics.
- Zeyu Huang, Yikang Shen, Xiaofeng Zhang, Jie Zhou, Wenge Rong, and Zhang Xiong. 2023. [Transformer-patcher: One mistake worth one neuron](#). In *The Eleventh International Conference on Learning Representations*.
- Inigo Jauregi Unanue, Najmeh Sadoughi, Vimal Bhat, Zhu Liu, and Massimo Piccardi. 2026. [Attribute-controlled translation with preference optimization](#). In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 4031–4057, Rabat, Morocco. Association for Computational Linguistics.
- Houcheng Jiang, Junfeng Fang, Ningyu Zhang, Mingyang Wan, Guojun Ma, Xiang Wang, Xiangnan He, and Tat-Seng Chua. 2025. [Anyedit: Edit any knowledge encoded in language models](#). In *Forty-second International Conference on Machine Learning*.
- Zehui Jiang, Xin Zhao, Yuta Kumadaki, and Naoki Yoshinaga. 2026. [CLICKER: Cross-lingual knowledge editing via in-context learning with adaptive stepwise reasoning](#). In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 5007–5022, Rabat, Morocco. Association for Computational Linguistics.
- Urvashi Khandelwal, Angela Fan, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. 2021. [Nearest neighbor machine translation](#). In *International Conference on Learning Representations*.
- Shuang Li, Jiangjie Chen, Siyu Yuan, Xinyi Wu, Hao Yang, Shimin Tao, and Yanghua Xiao. 2024. [Translate meanings, not just words: Idiomkb’s role in optimizing idiomatic translation with language models](#). *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):18554–18563.
- Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. 2022. [Locating and editing factual associations in GPT](#). In *Advances in Neural Information Processing Systems*.
- Kevin Meng, Arnab Sen Sharma, Alex J Andonian, Yonatan Belinkov, and David Bau. 2023. [Mass-editing memory in a transformer](#). In *The Eleventh International Conference on Learning Representations*.
- Eric Mitchell, Charles Lin, Antoine Bosselut, Chelsea Finn, and Christopher D Manning. 2022a. [Fast model editing at scale](#). In *International Conference on Learning Representations*.
- Eric Mitchell, Charles Lin, Antoine Bosselut, Christopher D Manning, and Chelsea Finn. 2022b. [Memory-based model editing at scale](#). In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 15817–15831. PMLR.
- Masatoshi Otake and Yusuke Miyao. 2023. [Comprehensive evaluation of translation error correction models](#). In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 862–871, Hong Kong, China. Association for Computational Linguistics.
- Stefano Perrella, Lorenzo Proietti, Pere-Lluís Hugué Cabot, Edoardo Barba, and Roberto Navigli. 2024. [Beyond correlation: Interpretable evaluation of machine translation metrics](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20689–20714, Miami, Florida, USA. Association for Computational Linguistics.
- Chenmian Tan, Ge Zhang, and Jie Fu. 2024. [Massive editing for large language models via meta learning](#). In *The Twelfth International Conference on Learning Representations*.
- Jiaan Wang, Yunlong Liang, Zengkui Sun, Yuxuan Cao, Jiarong Xu, and Fandong Meng. 2024a. [Cross-lingual knowledge editing in large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11676–11686, Bangkok, Thailand. Association for Computational Linguistics.
- Peng Wang, Zexi Li, Ningyu Zhang, Ziwen Xu, Yunzhi Yao, Yong Jiang, Pengjun Xie, Fei Huang, and Hua-jun Chen. 2024b. [WISE: Rethinking the knowledge memory for lifelong model editing of large language models](#). In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.
- Peng Wang, Ningyu Zhang, Bozhong Tian, Zekun Xi, Yunzhi Yao, Ziwen Xu, Mengru Wang, Shengyu Mao, Xiaohan Wang, Siyuan Cheng, Kangwei Liu, Yuansheng Ni, Guozhou Zheng, and Huajun Chen. 2024c. [EasyEdit: An easy-to-use knowledge editing framework for large language models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 82–93, Bangkok, Thailand. Association for Computational Linguistics.
- Zihao Wei, Jingcheng Deng, Liang Pang, Hanxing Ding, Huawei Shen, and Xueqi Cheng. 2025. [MLaKE: Multilingual knowledge editing benchmark for large language models](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4457–4473, Abu Dhabi, UAE. Association for Computational Linguistics.
- Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. [A paradigm shift in machine translation: Boosting translation performance of large language models](#). In *The Twelfth International Conference on Learning Representations*.

- Jianhao Yan, Pingchuan Yan, Yulong Chen, Jing Li, Xi-anchao Zhu, and Yue Zhang. 2024. Benchmarking gpt-4 against human translators: A comprehensive evaluation across languages, domains, and expertise levels. *arXiv preprint arXiv:2411.13775*.
- Zhengdong Yang, Sheng Li, and Chenhui Chu. 2025. [Generative error correction for emotion-aware speech-to-text translation](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 20413–20421, Vienna, Austria. Association for Computational Linguistics.
- Zexuan Zhong, Zhengxuan Wu, Christopher Manning, Christopher Potts, and Danqi Chen. 2023. [MQuAKE: Assessing knowledge editing in language models via multi-hop questions](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15686–15702, Singapore. Association for Computational Linguistics.
- Jianing Zhou, Hongyu Gong, and Suma Bhat. 2021. [PIE: A parallel idiomatic expression corpus for idiomatic sentence generation and paraphrasing](#). In *Proceedings of the 17th Workshop on Multiword Expressions (MWE 2021)*, pages 33–48, Online. Association for Computational Linguistics.
- Dawei Zhu, Pinzhen Chen, Miaoran Zhang, Barry Haddow, Xiaoyu Shen, and Dietrich Klakow. 2024a. [Fine-tuning large language models to translate: Will a touch of noisy data in misaligned languages suffice?](#) In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 388–409, Miami, Florida, USA. Association for Computational Linguistics.
- Dawei Zhu, Sony Trenous, Xiaoyu Shen, Dietrich Klakow, Bill Byrne, and Eva Hasler. 2024b. [A preference-driven paradigm for enhanced translation with large language models](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3385–3403, Mexico City, Mexico. Association for Computational Linguistics.

A Prompt Used

Prompts used in benchmark construction are provided in Table 5, 6 and 7. Translation prompt is provided in Table 7. Prompt for idiom usage verification is provided in Table 8. Translation evaluation prompt (ClaudeBi and ClaudeScale) is provided in Table 9. Prompt for constructing literal negative samples is provided in Table 10.

B Benchmark Design Clarifications

Automatic construction. MEMT-bench is designed to evaluate model editing, not to rank MT systems by absolute translation quality. The central signal is whether an editing method moves a model from an erroneous behavior toward a controlled target while preserving unrelated behavior. This relative setup follows the role of automatically constructed data in common editing benchmarks such as ZsRE (De Cao et al., 2021) and CounterFact (Meng et al., 2022), where the emphasis is editability and locality rather than human preference over all valid outputs.

Why the editing target need not be a unique human-optimal translation. As defined in Section 3, reliability and generalization measure whether the post-edit output scores higher than the pre-edit output on the same sample against the same reference, while locality measures whether the post-edit output is not degraded under the same conditions. These metrics therefore quantify a pre/post-edit change with respect to a fixed editing target, rather than proximity to a unique gold translation. In principle, any fixed reference would yield a well-defined measurement; in our setting, however, we deliberately select realistic, meaning-preserving translations whose distribution is close to the model’s original outputs. This makes the task more realistic and challenging, because measurable improvement requires a targeted edit rather than a coarse distributional shift. We do not claim that each target is the unique optimal translation, but our quality-control procedures ensure that each target is a valid translation suitable for editing. This distinction is also consistent with CounterFact (Meng et al., 2022), whose targets are deliberately false statements: the benchmark evaluates movement toward a specified target, not whether that target is universally true.

Human verification of reference quality. To directly assess the quality of the editing targets, we

Direction	# Samples	Acceptable Translation	Idiom Meaning Correct	Strict Pass	Avg. Score
EN→ZH	50	94.0%	98.0%	94.0%	4.76
EN→AR	50	98.0%	98.0%	98.0%	4.74
Overall	100	96.0%	98.0%	96.0%	4.75

Table 4: Human verification of reference quality. Strict Pass requires both an acceptable full-sentence translation and correct idiomatic meaning.

conducted a human verification study on EN→ZH and EN→AR. We selected EN→ZH because it is the main in-pair direction in our single-edit setting, and EN→AR because it tests transfer to a typologically and script-wise distant target language under the same English-source edits. We sampled 100 references and used a conservative *Strict Pass* criterion: the full sentence had to be an acceptable translation and convey the idiomatic meaning correctly. Annotators also assigned an overall quality score from 1 to 5. The results are shown in Table 4; the high strict-pass rates directly support the quality of the constructed references, especially on the idiom-semantics dimension.

References and paraphrases. Plain paraphrases are used as meaning-preserving pivots during reference construction. The final evaluation compares model outputs with target-language references, not with the plain paraphrases themselves. Thus, the benchmark does not assume that plain wording is the only acceptable translation; multi-reference extensions with idiomatic target-language variants would be complementary future work.

Editing descriptors. An editing descriptor specifies a source input, target language, and desired translation for a model behavior that has already been identified as erroneous. Descriptors are therefore controlled behavioral targets for evaluating editing algorithms. They are not intended to serve as general-purpose training data for improving arbitrary third-party MT systems.

Metric use. During descriptor construction, we combine COMET, BLEURT, MetricX, LLM judgments, and contrastive literal checks to select high-confidence idiom translation errors. During editing evaluation, we report paired pre/post-edit changes with BLEURT for efficiency and consistency. Since the comparison is within the same sample before and after editing, the relative change is the key signal.

Batch and sequential editing. Suppose three descriptors correct sentences A, B, and C. Sequen-

tial editing first edits the model for A, then edits the updated model for B, and finally edits the updated model for C. Batch editing instead applies one update intended to satisfy A, B, and C simultaneously. Both settings occur during editing, before inference.

C Bench Existing Translators on IdiomMT

We evaluate two specialized translation models (M2M100-1.2B and NLLB-200-3.3B) and two instruction-tuned LLMs (QWEN2.5-INSTRUCT-3B and QWEN2.5-INSTRUCT-7B) on the idiomMT-en2x/zh2x dataset. Detailed results are reported in Table 11, with key findings summarized below.

For English-source translations (en→x), NLLB-200-3.3B achieves the highest scores for English idioms translated into German, French, Japanese, and Arabic. However, QWEN2.5-INSTRUCT-7B consistently outperforms all other models in the English-to-Chinese direction.

For Chinese-source translations (zh→x), QWEN2.5-INSTRUCT-7B leads in most directions (Chinese→English, German, French, and Japanese), with QWEN2.5-INSTRUCT-3B as a strong second. In contrast, NLLB-200-3.3B slightly outperforms others on Chinese-to-Arabic translation.

Across both source settings, Arabic consistently emerges as the most challenging target language, with all models showing markedly lower performance compared to other targets.

Overall, these results indicate that idiom translation remains a challenging open problem, leaving ample room for future improvement.

Practical Recommendations Based on the observed patterns, we recommend using QWEN2.5-INSTRUCT-7B or QWEN2.5-INSTRUCT-3B for translations involving Chinese-either as source or as target (except when the target is Arabic). For translations from English into European languages (German, French) or Arabic, as well as from Chinese into Arabic, NLLB-200-3.3B remains the preferred choice.

D Causal Tracing Analysis

The full causal tracing results across each directions are visualized in both heatmaps (Figure 5 and 6) and line graphs (Figure 7 and 8).

E Editing Methods

FT It updates parameters by gradient descent for a single MLP layers. We adopt FT-M in EasyEdit (Wang et al., 2024c) which finetune a single layer by cross-entropy loss optimization.

ROME (Meng et al., 2022) The first locate-then-edit editing method. It is proposed to insert or modify a single factual knowledge without retraining the model. ROME identifies a key vector k^* , which is the hidden activation in an MLP layer when the subject token appears (in our study, we use the last idiom token according to the *causal tracing analysis*). It then computes a value vector v^* representing the desired output representation. The layer weight matrix W is then updated by a rank-one modification: $\widehat{W} = W + (v^* - Wk^*)$. The rank-one modification ensures that the update changes the model’s behavior only along one semantic direction (k^*), preserving the rest of the model’s knowledge.

MEMIT (Meng et al., 2023) It generalizes ROME to edit multiple facts simultaneously. MEMIT identifies causal MLP layers that mediate factual recall and computes key-value pairs for each target $\{(k_i, v_i)\}_{i=1}^N$. It enforces $\Delta W k_i = v_i - W k_i$ for all targets, which can be compactly expressed as $\Delta W K = V_\Delta$, with K and V_Δ denoting the stacked key matrix and value difference matrix, respectively. The resulting update is obtained in closed form as $\Delta W = V_\Delta K^+$, where the Moore-Penrose pseudo-inverse K^+ provides the minimum-norm solution to the linear constraints.

AlphaEdit (Fang et al., 2025) AlphaEdit seeks to preserve existing knowledge while integrating new information. Naive updates can distort unrelated representations, so AlphaEdit introduces a constraint mechanism. After identifying the editable weights, AlphaEdit computes an update direction (similar to $v^* - Wk^*$) that encodes the new knowledge. It then projects this update onto the null space of features representing preserved knowledge (we use [multilingual 20231101-Wikipedia](#) data to estimate the preserved knowledge matrix). Formally, if P denotes the projection matrix onto the null space of preserved features, the final update is given by $\Delta W = P(v^* - Wk^*)k^\top$. This null-space projection ensures that the edit does not alter directions associated with other important facts.

UNKE (Deng et al., 2025) It extends model editing from structured to unstructured knowledge, where the knowledge cannot be represented in a clear (subject, relation, object) format. UNKE replaces the local MLP key-value assumption with a non-local block key-value mechanism that integrates information from multiple transformer layers. Instead of locating and editing specific subject tokens, it performs cause-driven optimization by directly adjusting the hidden representation of the final token, which summarizes the full input context.

GRACE (Hartvigsen et al., 2023) GRACE is proposed to enable many sequential edits without overwriting the base model parameters. Rather than modifying the transformer’s weights, GRACE attaches adaptor modules containing discrete key-value codebooks at specific layers. During inference, when a new prompt is fed into the model, GRACE computes its latent representation at the adaptor layer and performs nearest-key retrieval from the codebook. If the prompt’s latent vector is close to a stored key (indicating that it is related to a previously edited fact), the corresponding transformation (the “edit”) is applied to the layer output. If no key in the codebook is sufficiently similar which means that the input is unrelated to any prior edit, GRACE simply passes through the original layer output unchanged.

WISE (Wang et al., 2024b) WISE aims to balance editability and model integrity by isolating edited parameters while maintaining shared inference pathways. The WISE framework partitions the model’s knowledge memory into two distinct parametric modules: a *main memory* retaining the original pretrained parameters and a *side memory* dedicated solely to edited knowledge. During edits, only the side memory is modified, while the main memory remains untouched. A dedicated router module is trained via supervised signals that contrast edited-scope inputs with non-edited inputs (randomly sampled from OPUS-100), enabling dynamic routing of queries at inference time either through the main memory or the side memory.

F Model Editing Results

The detailed results on reliability, generalization, our locality set, flores, and MMLU test for English to Chinese editing descriptor are provided in Table 12, 13, 14, 15, and 16. The detailed results

on reliability, generalization, our locality set, flores, and MMLU test for Chinese to English editing descriptor are provided in Table 17, 18, 19, 20, and 21.

G Examples

G.1 IdiomMT Instance Examples

Table 22 provides an instance example in dataset IdiomMT-en2x. Table 23 provides an instance example in dataset IdiomMT-zh2x.

G.2 Editing Descriptors Examples

Table 24 provides an editing descriptor example from Translator Qwen2.5-3B-Instruct (source sentence in English). Table 25 provides an editing descriptor example from Translator Qwen2.5-3B-Instruct (source sentence in Chinese).

H Case Analysis

Table 26 and Table 27 show good case of model editing on translation. Table 28, Table 29, Table 33, Table 34, Table 32, Table 31 and Table 30 provides bad cases. Figure 13 shows the language shift issues in large batch size and sequential editing.

Prompt Used In Step 1&2: EN Idiom → (EN Idiomatic Sent, EN Plain Sent)
You are tasked with generating structured outputs for English idioms. For each given idiom and its meaning in English, follow these steps: 1. English Sentence with Idiom - Create an extremely short sentence (≤ 15 words) in English that uses the idiom naturally and correctly. - The sentence must be clear, self-contained, and unambiguous. - The sentence must be concise and must not add extra context or redundant explanation beyond the idiom's meaning. 2. Plain English Sentence - Rewrite the idiom sentence by replacing the idiom with a plain, direct expression of its meaning. - Make only minimal changes to the sentence structure. Keep wording and grammar as close to the idiom sentence as possible. - Do not introduce new context or elaboration. Output Format: Return a valid JSON object with the following keys only: <pre>{ "English": "<English sentence with idiom>", "Plain_English": "<Plain English sentence>" }</pre> English Idiom: {Idiom} English Meaning: {Meaning}

Table 5: Prompts used in the step 1 and 2 of IdiomMT-en2x construction process

Prompt Used In Step 1&2&3: ZH Idiom → (ZH Idiomatic Sent, ZH Plain Sent, EN Trans)
You are tasked with generating structured outputs for Chinese idioms. For each given idiom and its meaning in Chinese and English, follow these steps: 1. Chinese Sentence with Idiom - Create a extreme short sentence (≤ 15 words) in Chinese that uses the idiom naturally and correctly. - The sentence must be clear, self-contained, and unambiguous. - The sentence must be concise and must not add extra context or redundant explanation beyond the idiom's meaning. 2. Plain Chinese Sentence - Rewrite the idiom sentence by replacing the idiom with a plain, direct expression of its meaning. - Make only minimal changes to the sentence structure. Keep wording and grammar as close to the idiom sentence as possible. - Do not introduce new context or elaboration. 3. Plain English Translation - Based on the parallel Chinese sentence and the given English meaning, provide a clear, simple English sentence. - The English must reflect the idiomatic meaning, without using any idiom. Output Format: - Return a valid JSON object with the following keys only: <pre>{ "Chinese": "<chinese sentence with idiom>", "Plain_Chinese": "Plain Chinese Sentence", "English": "<plain English translation>" }</pre> Chinese Idiom: {Idiom} Chinese Meaning: {Meaning} English Meaning: {Meaning}

Table 6: Prompts used in the step 1 ,2 and 3 of IdiomMT-zh2x construction process

Prompt Used For General Translation: Source language $\rightarrow$ Target Language

Translate text from {SLang} to {TLang}.

- The translation must be accurate, fluent, and preserve the original meaning.
- Do not add explanations, notes, or extra content.
- Output only the translated text.

{SLang}: {Source}

{TLang}:

Table 7: Prompts used for general translation. It used in the step 3 of IdiomMT-en2x and IdiomMT-zh2x construction process, and used in prompting QWEN2.5-INSTRUCT as translator.

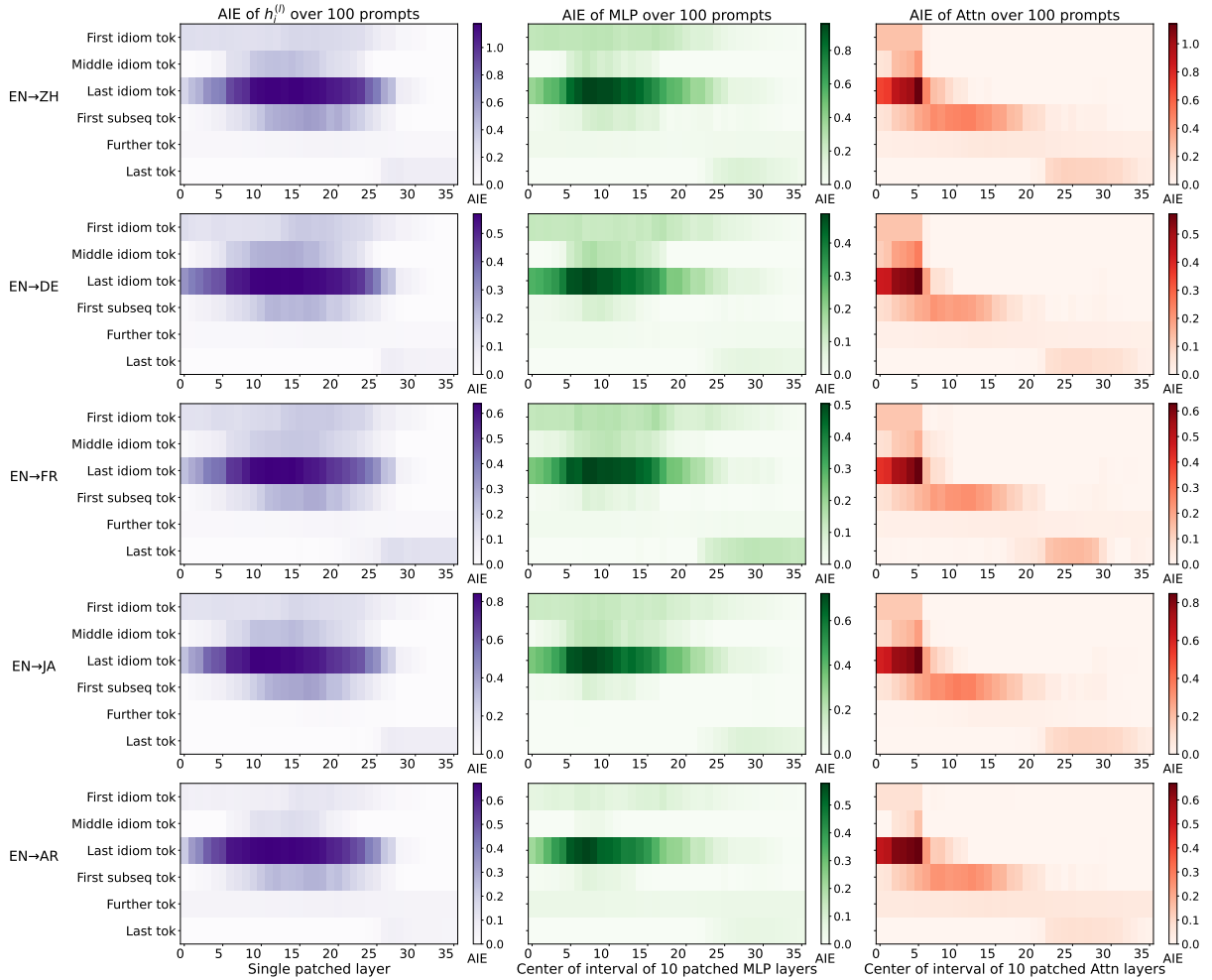


Figure 5: Mean causal traces of Qwen2.5-3B-Instruct over a sample of 100 English $\rightarrow$ X translation, shown as a heatmaps.

<u>Prompt Used In English Idiom Detection</u> You are given an idiom (sometimes written with blanks to indicate variable words) and a sentence that contains that idiom. Your task is to identify and output the exact span of the idiom as it appears in the sentence. - Match idioms even if they appear in a different tense, form, or with small inserted words. - Only output the idiom span exactly as it occurs in the sentence. - Do not output explanations or extra text. EXAMPLE 1: Input: Idiom: hit the _ Sentence: After the long trip, he was so tired that he hit the bed immediately. Output: hit the bed EXAMPLE 2: Input: Idiom: let the cat out of the bag Sentence: She accidentally let the cats out of the bag during the conversation. Output: let the cats out of the bag EXAMPLE 3: Input: Idiom: throw in the towel Sentence: Realizing the competition was too strong, he finally threw in the dirty towel. Output: threw in the dirty towel Input: Idiom: {Idiom} Sentence: {Sentence} Output:
<u>Prompt Used In Chinese Idiom Detection</u> You are given a Chinese idiom and a Chinese sentence that contains that idiom. Your task is to identify and output the exact span of the idiom as it appears in the sentence. - When matching, treat simplified and traditional characters as equivalent. - Only output the idiom span exactly as it occurs in the sentence. - Do not output explanations or extra text. EXAMPLE 1: Input: Idiom: 随心所欲 Sentence: 他做事總是随心所欲，從不顧及他人感受。 Output: 随心所欲 Input: Idiom: {Idiom} Sentence: {Sentence} Output:

Table 8: Prompts used in idiom usage verification.

Claude-Scale Translation Evaluation Prompt Score the following {SLang} to {TLang} translation with respect to the human reference on a continuous scale from 0 to 100, where score of 0 means "no meaning preserved" and score of 100 means "perfect meaning and grammar". {TLang} human reference: {Reference} {TLang} translation: {Translation} Score (0-100):
Claude-Bi Translation Evaluation Prompt Score the following {SLang} to {TLang} translation with respect to the human reference on a binary scale, where a score of 0 means "unacceptable" and a score of 1 means "acceptable". {TLang} human reference: {Reference} {TLang} translation: {Translation} Score (0 or 1):

Table 9: Translation Evaluation Prompt

Prompt Used In Literal Negative Translation Produce a bad example by generating an unnatural, strictly literal word-for-word translation from {SLang} to {TLang}. - The translation must deliberately reflect a word-for-word mapping, ignoring idioms and natural fluency. - Do not add explanations, notes, or extra content. - Output only the literal translation text. {SLang}: {Source} {TLang}:

Table 10: Prompt for constructing literal negative samples

	English-Chinese				English-Japanese				English-German				English-French				English-Arabic			
	m2m100	nllb	qwen3b	qwen7b	m2m100	nllb	qwen3b	qwen7b	m2m100	nllb	qwen3b	qwen7b	m2m100	nllb	qwen3b	qwen7b	m2m100	nllb	qwen3b	qwen7b
COMET (↑)	0.688	0.759	0.810	0.826	0.762	0.827	0.807	0.816	0.682	0.766	0.716	0.742	0.680	0.761	0.714	0.749	0.643	0.776	0.674	0.731
BLEURT-20 (↑)	0.503	0.573	0.616	0.639	0.493	0.521	0.539	0.551	0.545	0.654	0.597	0.624	0.478	0.594	0.518	0.581	0.461	0.578	0.438	0.537
MetricX-23 (↓)	3.166	2.367	1.501	1.398	2.742	1.849	1.799	1.925	1.839	1.174	1.355	1.278	2.463	1.559	1.948	1.637	4.153	2.015	2.662	2.627
Claude-Scale (↑)	40.280	56.471	70.309	74.598	37.963	56.241	51.659	56.097	53.316	69.820	57.758	64.579	57.137	71.621	61.007	70.731	27.237	63.525	38.149	55.511
Claude-Binary (↑)	0.136	0.304	0.533	0.616	0.111	0.306	0.243	0.323	0.247	0.503	0.277	0.374	0.284	0.533	0.330	0.495	0.048	0.396	0.170	0.340
	Chinese-English				Chinese-Japanese				Chinese-German				Chinese-French				Chinese-Arabic			
	m2m100	nllb	qwen3b	qwen7b	m2m100	nllb	qwen3b	qwen7b	m2m100	nllb	qwen3b	qwen7b	m2m100	nllb	qwen3b	qwen7b	m2m100	nllb	qwen3b	qwen7b
COMET (↑)	0.729	0.764	0.810	0.817	0.800	0.805	0.838	0.838	0.702	0.744	0.750	0.775	0.696	0.736	0.744	0.781	0.624	0.750	0.639	0.684
BLEURT-20 (↑)	0.616	0.665	0.704	0.715	0.569	0.528	0.593	0.601	0.609	0.665	0.650	0.684	0.523	0.586	0.568	0.641	0.458	0.598	0.337	0.465
MetricX-23 (↓)	2.249	1.625	1.412	1.439	1.689	1.660	1.108	1.355	1.220	0.948	0.932	0.908	1.618	1.127	1.214	1.040	3.271	1.317	2.100	2.667
Claude-Scale (↑)	49.000	60.886	76.940	78.916	42.961	46.915	59.147	60.468	40.990	50.830	58.557	65.935	44.409	54.288	62.222	71.169	23.056	47.880	28.889	45.418
Claude-Binary (↑)	0.206	0.407	0.687	0.730	0.120	0.172	0.313	0.339	0.104	0.211	0.272	0.384	0.127	0.255	0.350	0.494	0.029	0.199	0.107	0.258

Table 11: Translation Performance across language pairs on IdiomMT-en2x/zh2x

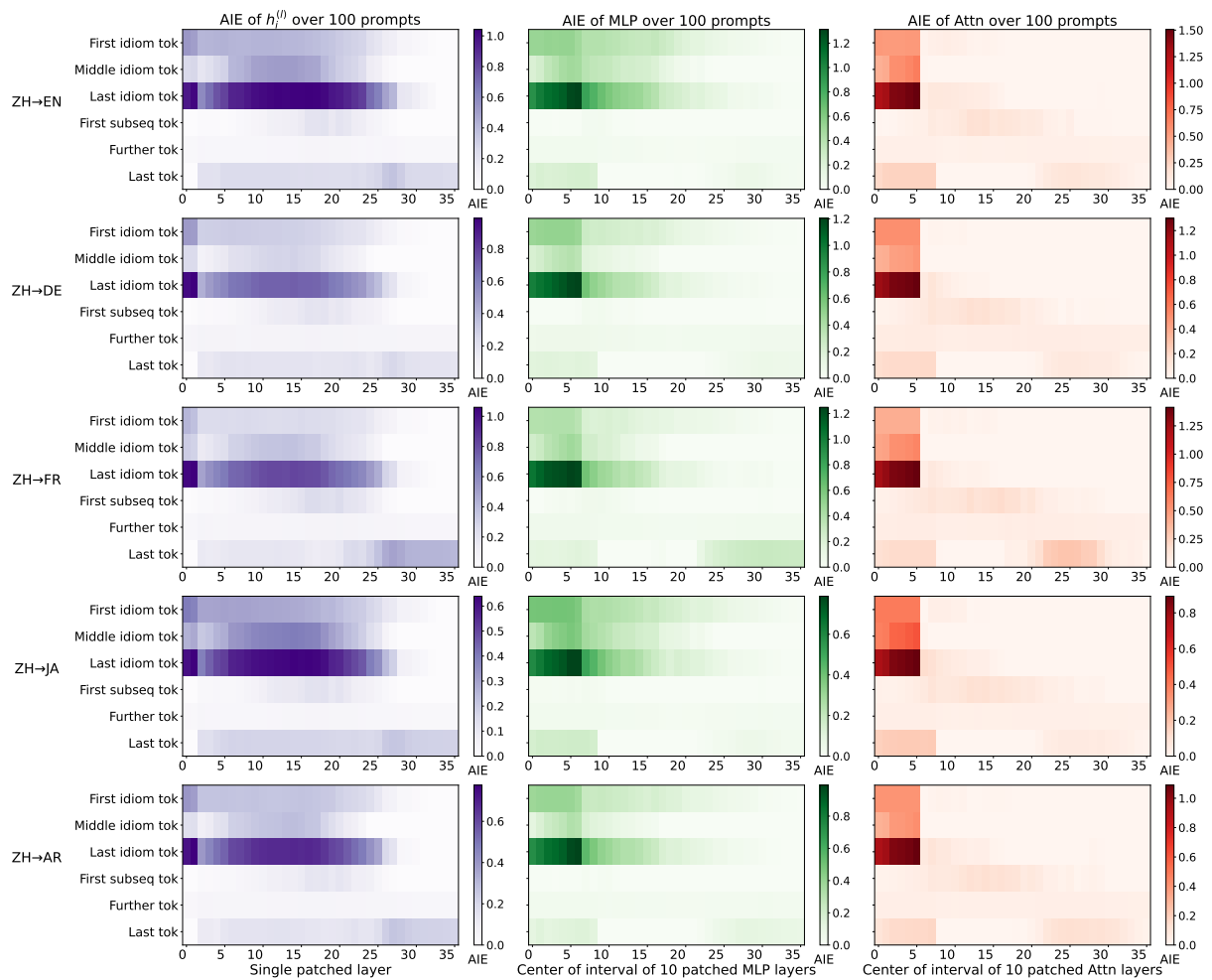


Figure 6: Mean causal traces of Qwen2.5-3B-Instruct over a sample of 100 Chinese→X translation, shown as a heatmaps.

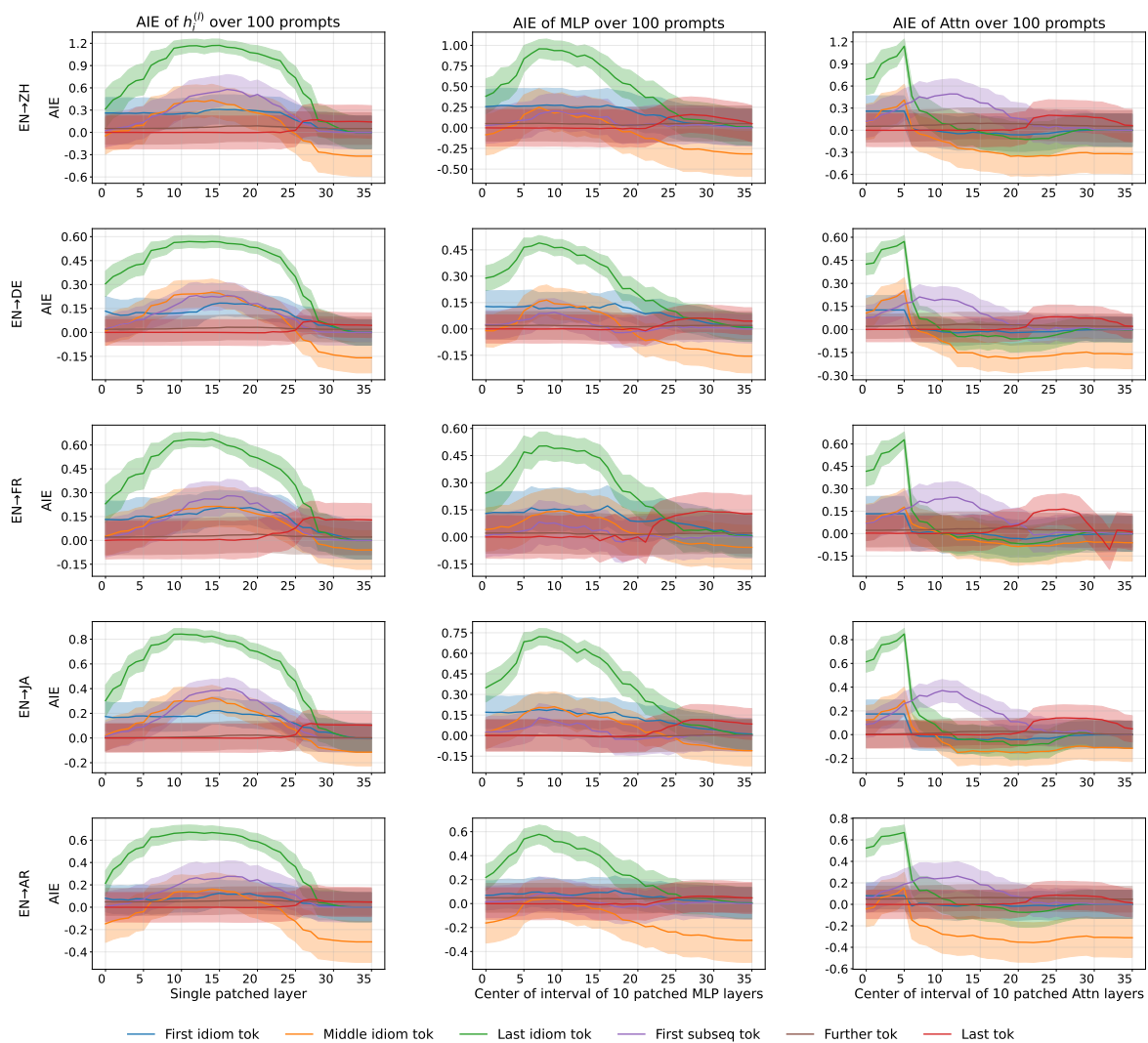


Figure 7: Mean causal traces of Qwen2.5-3B-Instruct over a sample of 100 English→X translation, shown as line plots with 95% confidence intervals.

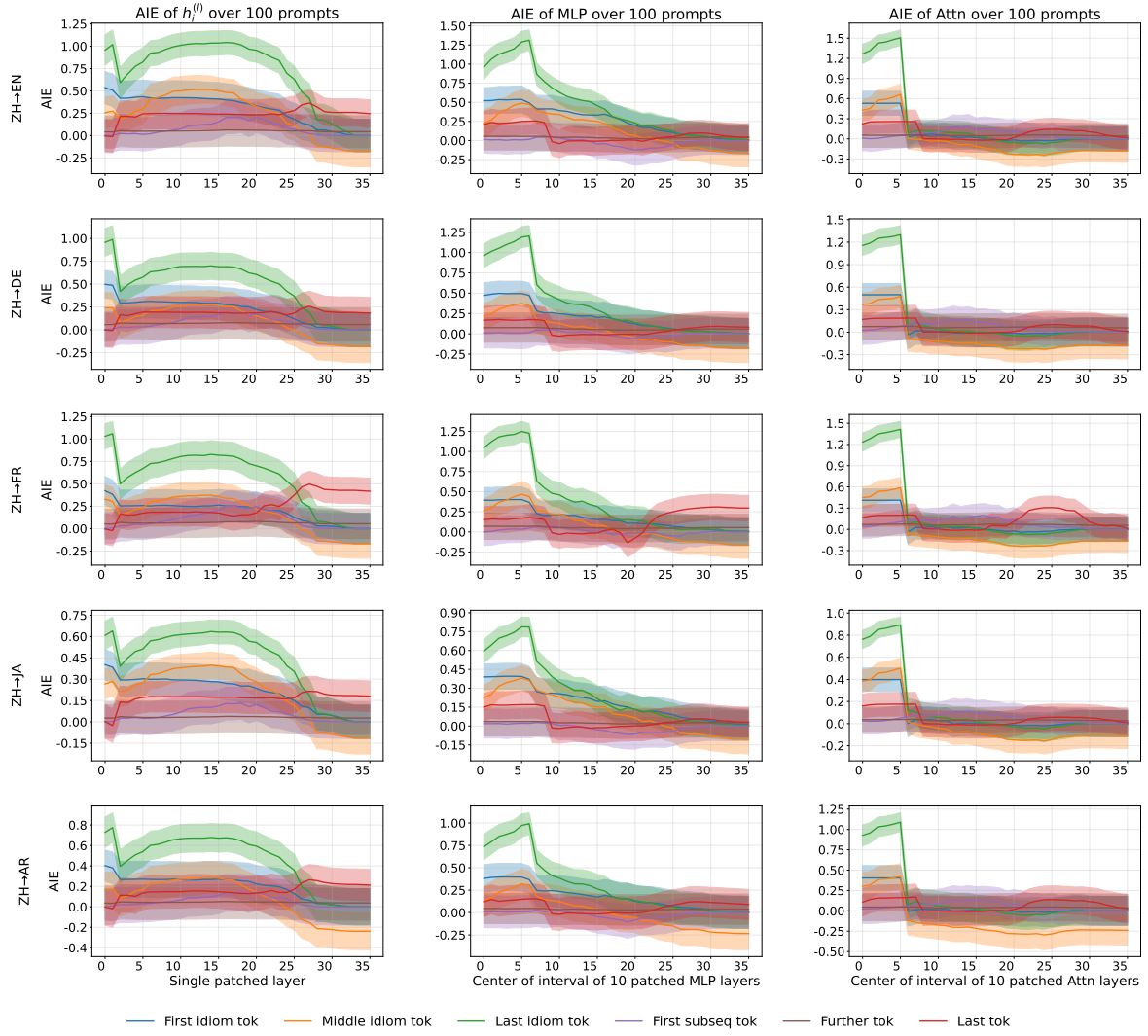


Figure 8: Mean causal traces of Qwen2.5-3B-Instruct over a sample of 100 Chinese→X translation, shown as line plots with 95% confidence intervals.

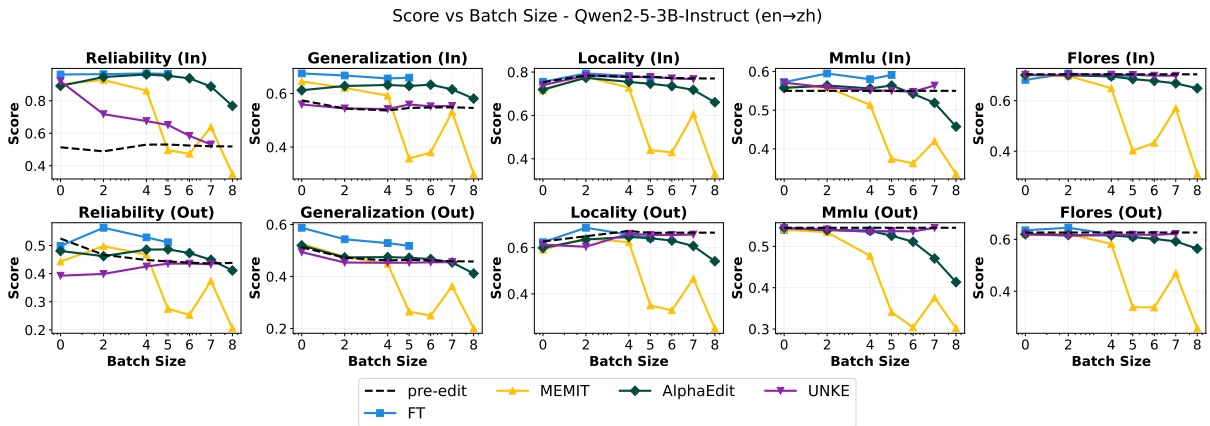


Figure 9: Editing performance versus batch size for English-to-Chinese descriptor editing. The x-axis shows batch sizes as powers of two. Due to memory constraints on a single 40GB GPU, FT supports batch sizes up to 2^5 , while UNKE supports up to 2^7 .

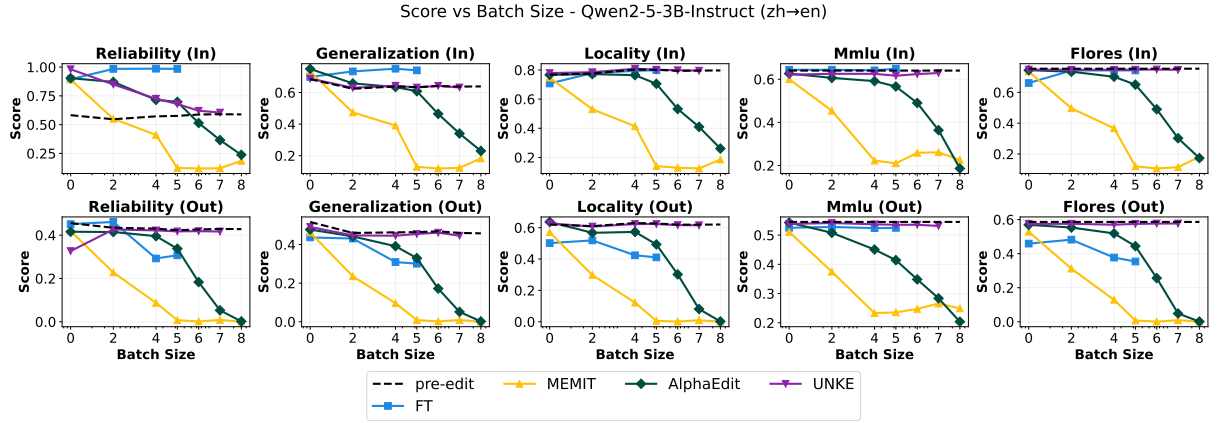


Figure 10: Editing performance versus batch size for Chinese-to-English descriptor editing. The x-axis shows batch sizes as powers of two. Due to memory constraints on a single 40GB GPU, FT supports batch sizes up to 2^5 , while UNKE supports up to 2^7 .

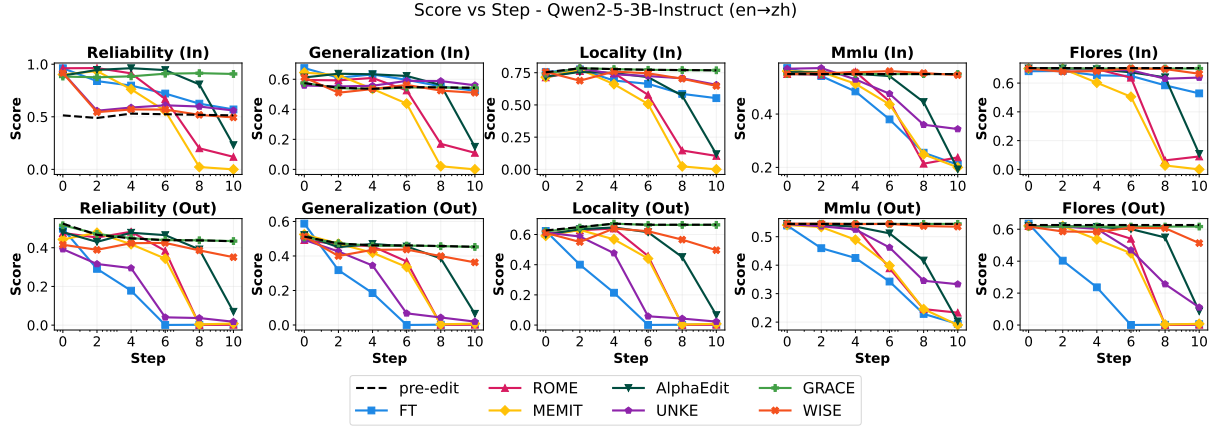


Figure 11: Editing performance versus sequential step for English-to-Chinese descriptor editing. The x-axis represents batch sizes in powers of two.

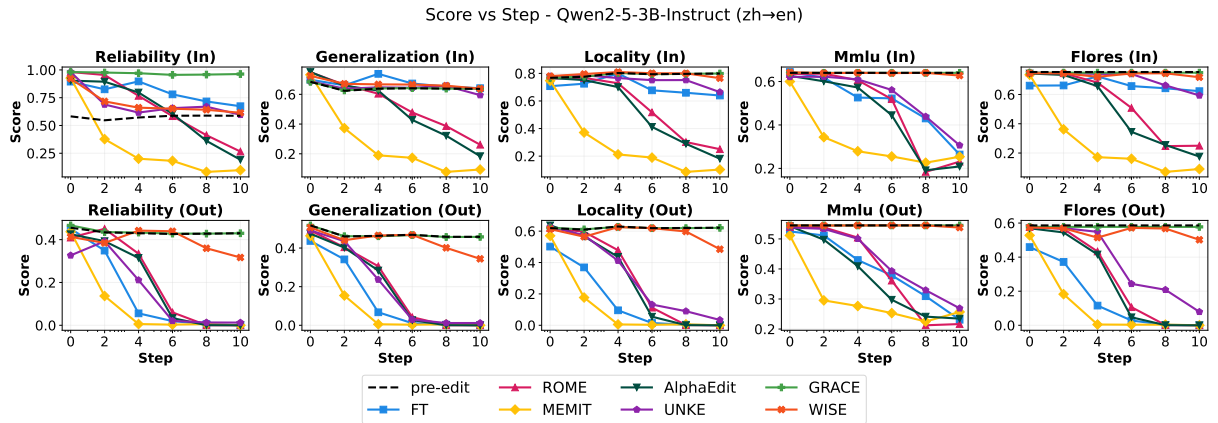
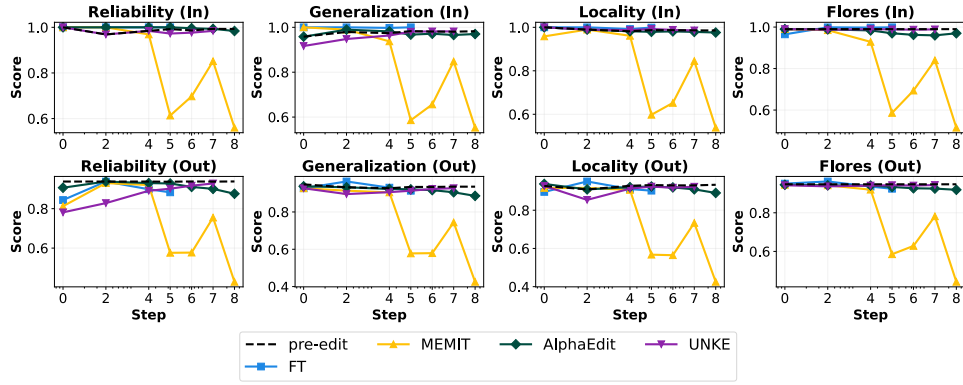
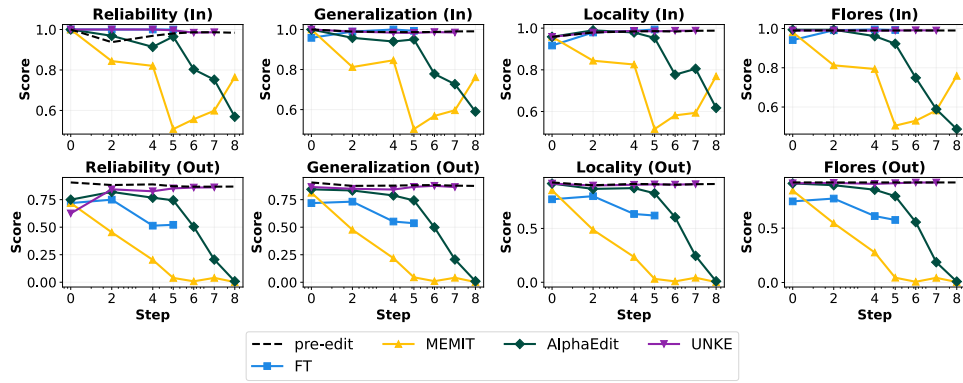


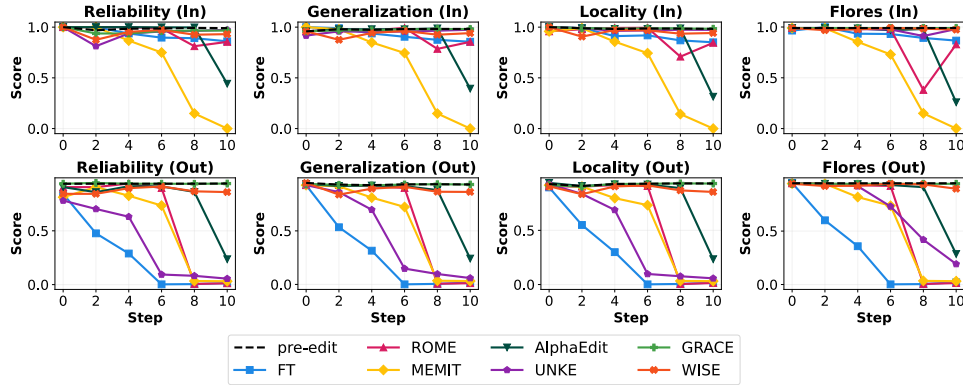
Figure 12: Editing performance versus sequential step for Chinese-to-English descriptor editing. The x-axis represents batch sizes in powers of two.



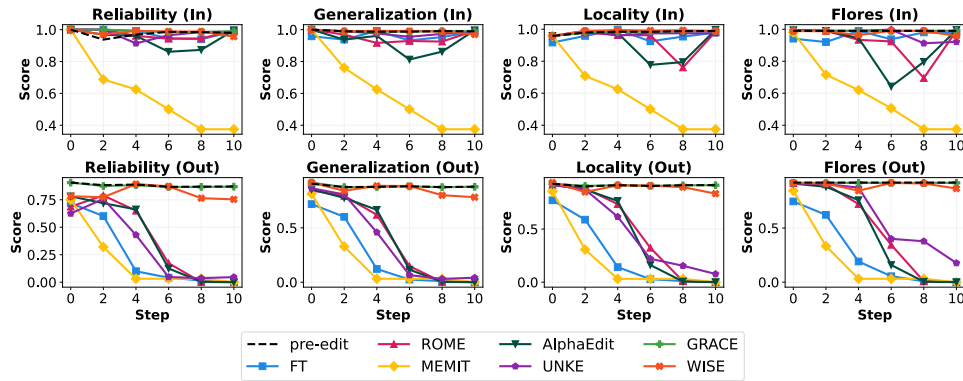
(a) Batch Editing (Descriptor: English to Chinese)



(b) Batch Editing (Descriptor: Chinese to English)



(c) Sequential Editing (Descriptor: English to Chinese)



(d) Sequential Editing (Descriptor: Chinese to English)

Figure 13: Language detection results in batch editing and sequential editing.

	Chinese		German		French		Japanese		Arabic	
	Score	Impr.	Score	Impr.	Score	Impr.	Score	Impr.	Score	Impr.
pre-edit	0.511	-	0.515	-	0.414	-	0.455	-	0.351	-
FT	0.954	0.987	0.585	0.630	0.525	0.668	0.479	0.596	0.419	0.582
ROME	0.961	0.999	0.543	0.561	0.456	0.529	0.480	0.547	0.380	0.507
MEMIT	0.938	0.987	0.564	0.580	0.489	0.613	0.507	0.614	0.386	0.517
AlphaEdit	0.963	0.998	0.559	0.577	0.483	0.600	0.478	0.555	0.379	0.511
UNKE	0.942	0.982	0.444	0.400	0.336	0.368	0.394	0.432	0.287	0.328
GRACE	0.917	0.961	0.516	0.108	0.414	0.043	0.455	0.089	0.350	0.091
WISE	0.840	0.816	0.512	0.303	0.407	0.272	0.440	0.332	0.336	0.166

Table 12: **Reliability** performance comparison of different editing methods tested on each target language when editing descriptor is **English to Chinese**.

	Chinese		German		French		Japanese		Arabic	
	Score	Impr.	Score	Impr.	Score	Impr.	Score	Impr.	Score	Impr.
pre-edit	0.542	-	0.548	-	0.456	-	0.470	-	0.354	-
FT	0.651	0.706	0.591	0.593	0.531	0.624	0.521	0.611	0.445	0.599
ROME	0.625	0.654	0.557	0.500	0.466	0.500	0.485	0.524	0.378	0.467
MEMIT	0.640	0.690	0.572	0.538	0.486	0.525	0.498	0.560	0.386	0.491
AlphaEdit	0.641	0.679	0.559	0.505	0.483	0.524	0.480	0.530	0.377	0.481
UNKE	0.557	0.320	0.532	0.292	0.430	0.224	0.451	0.288	0.336	0.254
GRACE	0.541	0.060	0.548	0.093	0.453	0.058	0.470	0.097	0.355	0.107
WISE	0.606	0.509	0.539	0.231	0.446	0.222	0.460	0.301	0.355	0.152

Table 13: **Generalization** performance comparison of different editing methods tested on each target language when editing descriptor is **English to Chinese**.

	Chinese		German		French		Japanese		Arabic	
	Score	Pres.	Score	Pres.	Score	Pres.	Score	Pres.	Score	Pres.
pre-edit	0.767	-	0.761	-	0.763	-	0.679	-	0.459	-
FT	0.749	0.527	0.744	0.595	0.744	0.623	0.626	0.475	0.475	0.591
ROME	0.756	0.769	0.750	0.773	0.746	0.809	0.668	0.781	0.455	0.808
MEMIT	0.756	0.766	0.754	0.783	0.749	0.823	0.671	0.796	0.456	0.807
AlphaEdit	0.752	0.741	0.751	0.769	0.741	0.795	0.664	0.767	0.456	0.804
UNKE	0.750	0.746	0.739	0.741	0.728	0.770	0.648	0.713	0.427	0.734
GRACE	0.768	0.905	0.762	0.922	0.763	0.956	0.678	0.929	0.458	0.923
WISE	0.754	0.731	0.750	0.787	0.753	0.812	0.658	0.722	0.456	0.892

Table 14: **Locality** performance comparison of different editing methods tested on each target language when editing descriptor is **English to Chinese**.

	Chinese		German		French		Japanese		Arabic	
	Score	Pres.	Score	Pres.	Score	Pres.	Score	Pres.	Score	Pres.
pre-edit	0.694	-	0.702	-	0.706	-	0.607	-	0.444	-
FT	0.682	0.496	0.686	0.534	0.694	0.538	0.607	0.556	0.465	0.586
ROME	0.694	0.857	0.701	0.851	0.704	0.871	0.608	0.797	0.444	0.779
MEMIT	0.694	0.862	0.699	0.837	0.705	0.874	0.609	0.798	0.447	0.767
AlphaEdit	0.694	0.850	0.700	0.844	0.704	0.871	0.609	0.793	0.449	0.791
UNKE	0.692	0.843	0.701	0.811	0.703	0.838	0.606	0.750	0.446	0.740
GRACE	0.694	0.888	0.699	0.877	0.704	0.899	0.610	0.829	0.446	0.817
WISE	0.689	0.775	0.692	0.782	0.697	0.798	0.604	0.702	0.439	0.771

Table 15: **Flores** performance comparison of different editing methods tested on each target language when editing descriptor is **English to Chinese**.

	English		Chinese		German		French		Japanese		Arabic	
	Score	Pres.	Score	Pres.	Score	Pres.	Score	Pres.	Score	Pres.	Score	Pres.
pre-edit	0.637	-	0.568	-	0.527	-	0.532	-	0.511	-	0.495	-
FT	0.639	0.965	0.571	0.961	0.524	0.953	0.549	0.953	0.502	0.947	0.494	0.939
ROME	0.633	0.995	0.570	0.998	0.532	0.994	0.556	0.995	0.510	0.994	0.492	0.995
MEMIT	0.636	0.995	0.568	0.993	0.530	0.988	0.557	0.994	0.507	0.988	0.494	0.989
AlphaEdit	0.635	0.992	0.569	0.995	0.533	0.992	0.556	0.993	0.513	0.995	0.492	0.994
UNKE	0.633	0.988	0.576	0.994	0.530	0.985	0.560	0.992	0.508	0.987	0.489	0.981
GRACE	0.637	1.000	0.568	1.000	0.527	1.000	0.552	1.000	0.511	1.000	0.495	1.000
WISE	0.637	0.992	0.563	0.989	0.526	0.991	0.561	0.993	0.507	0.986	0.497	0.991

Table 16: **MMLU** performance comparison of different editing methods tested on each target language when editing descriptor is **English to Chinese**.

	English		German		French		Japanese		Arabic	
	Score	Impr.	Score	Impr.	Score	Impr.	Score	Impr.	Score	Impr.
pre-edit	0.593	-	0.563	-	0.445	-	0.507	-	0.228	-
FT	0.971	0.983	0.632	0.673	0.470	0.565	0.442	0.483	0.294	0.471
ROME	0.980	0.997	0.558	0.508	0.475	0.554	0.445	0.415	0.185	0.311
MEMIT	0.885	0.899	0.506	0.476	0.423	0.489	0.427	0.404	0.179	0.301
AlphaEdit	0.936	0.970	0.552	0.511	0.458	0.528	0.466	0.434	0.204	0.350
UNKE	0.969	0.985	0.480	0.408	0.337	0.355	0.354	0.319	0.157	0.239
GRACE	0.966	0.982	0.562	0.100	0.442	0.095	0.508	0.088	0.231	0.092
WISE	0.805	0.757	0.518	0.310	0.422	0.367	0.458	0.314	0.223	0.138

Table 17: **Reliability** performance comparison of different editing methods tested on each target language when editing descriptor is **Chinese to English**.

	English		German		French		Japanese		Arabic	
	Score	Impr.	Score	Impr.	Score	Impr.	Score	Impr.	Score	Impr.
pre-edit	0.645	-	0.585	-	0.478	-	0.535	-	0.263	-
FT	0.722	0.743	0.593	0.561	0.455	0.492	0.508	0.509	0.303	0.461
ROME	0.701	0.660	0.541	0.432	0.451	0.456	0.483	0.416	0.222	0.321
MEMIT	0.658	0.587	0.493	0.378	0.409	0.404	0.458	0.403	0.205	0.298
AlphaEdit	0.689	0.624	0.538	0.419	0.437	0.441	0.491	0.440	0.223	0.333
UNKE	0.655	0.324	0.562	0.305	0.446	0.289	0.493	0.290	0.248	0.244
GRACE	0.645	0.079	0.585	0.112	0.477	0.100	0.535	0.098	0.264	0.090
WISE	0.674	0.470	0.536	0.262	0.431	0.272	0.492	0.292	0.260	0.119

Table 18: **Generalization** performance comparison of different editing methods tested on each target language when editing descriptor is **Chinese to English**.

	English		German		French		Japanese		Arabic	
	Score	Pres.	Score	Pres.	Score	Pres.	Score	Pres.	Score	Pres.
pre-edit	0.801	-	0.726	-	0.689	-	0.707	-	0.381	-
FT	0.763	0.552	0.688	0.526	0.577	0.432	0.631	0.456	0.346	0.566
ROME	0.796	0.775	0.721	0.742	0.679	0.781	0.701	0.753	0.374	0.794
MEMIT	0.754	0.659	0.657	0.627	0.618	0.639	0.642	0.631	0.328	0.673
AlphaEdit	0.781	0.631	0.706	0.604	0.653	0.598	0.688	0.597	0.373	0.671
UNKE	0.794	0.774	0.717	0.756	0.671	0.764	0.684	0.747	0.371	0.786
GRACE	0.800	0.860	0.726	0.854	0.690	0.924	0.707	0.867	0.380	0.919
WISE	0.781	0.717	0.667	0.628	0.637	0.672	0.653	0.615	0.374	0.890

Table 19: **Locality** performance comparison of different editing methods tested on each target language when editing descriptor is **Chinese to English**.

	English		German		French		Japanese		Arabic	
	Score	Pres.	Score	Pres.	Score	Pres.	Score	Pres.	Score	Pres.
pre-edit	0.744	-	0.667	-	0.614	-	0.629	-	0.329	-
FT	0.699	0.494	0.619	0.473	0.525	0.403	0.561	0.444	0.302	0.569
ROME	0.746	0.850	0.664	0.752	0.610	0.758	0.630	0.783	0.328	0.753
MEMIT	0.705	0.641	0.603	0.563	0.553	0.568	0.574	0.574	0.299	0.614
AlphaEdit	0.742	0.690	0.663	0.601	0.598	0.606	0.629	0.593	0.328	0.643
UNKE	0.744	0.858	0.665	0.741	0.610	0.783	0.629	0.749	0.340	0.763
GRACE	0.743	0.930	0.667	0.827	0.612	0.857	0.629	0.835	0.331	0.823
WISE	0.727	0.716	0.618	0.614	0.574	0.631	0.588	0.546	0.314	0.768

Table 20: **Flores** performance comparison of different editing methods tested on each target language when editing descriptor is **Chinese to English**.

	English		Chinese		German		French		Japanese		Arabic	
	Score	Pres.	Score	Pres.	Score	Pres.	Score	Pres.	Score	Pres.	Score	Pres.
pre-edit	0.637	-	0.568	-	0.527	-	0.552	-	0.511	-	0.495	-
FT	0.624	0.965	0.560	0.965	0.508	0.954	0.547	0.965	0.497	0.952	0.493	0.960
ROME	0.634	0.992	0.567	0.992	0.528	0.995	0.552	0.993	0.505	0.987	0.497	0.997
MEMIT	0.587	0.924	0.524	0.917	0.508	0.933	0.497	0.900	0.470	0.918	0.462	0.918
AlphaEdit	0.627	0.980	0.557	0.963	0.514	0.967	0.547	0.971	0.489	0.940	0.470	0.944
UNKE	0.639	0.992	0.563	0.992	0.516	0.982	0.555	0.986	0.500	0.977	0.488	0.980
GRACE	0.637	1.000	0.568	1.000	0.527	1.000	0.552	1.000	0.511	1.000	0.495	1.000
WISE	0.627	0.980	0.566	0.985	0.522	0.980	0.560	0.990	0.502	0.983	0.488	0.982

Table 21:

Idiom	1 o'clock sharp
Editing Descriptor (Reliability Test) Candidate	
EN Idiomatic-Sent.	The meeting starts at 1 o'clock sharp.
DE Reference	Das Meeting beginnt pünktlich um 1 Uhr.
FR Reference	La réunion commence ponctuellement à 1 heure.
ZH Reference	会议 1 点准时开始。
JA Reference	会議は 1 時きっかりに始まります。
AR Reference	يبدأ الاجتماع في تمام الساعة الواحدة.
Generalization Test	
Test 1:	
EN Source Sent.	Please arrive at 2 o'clock sharp for the interview.
DE Reference	Bitte erscheinen Sie pünktlich um 2 Uhr zum Vorstellungsgespräch.
FR Reference	Veuillez arriver ponctuellement à 2 heures pour l'entretien.
ZH Reference	请准时于 2 点到达参加面试。
JA Reference	面接のため 2 時にきちんと時間通りにお越しください。
AR Reference	يرجى الوصول في الوقت المحدد في الساعة الثانية للمقابلة.
Test 2:	
EN Source Sent.	The train departs at 9 o'clock sharp every morning.
DE Reference	Der Zug fährt jeden Morgen punkt 9 Uhr ab.
FR Reference	Le train part exactement à 9 heures tous les matins.
ZH Reference	火车每天早上 9 点整准时发车。
JA Reference	電車は毎朝 9 時きっかりに出発します。
AR Reference	القطار يغادر تماماً في الساعة التاسعة كل صباح.
Test 3:	
EN Source Sent.	The ceremony begins at noon sharp.
DE Reference	Die Zeremonie beginnt genau um zwölf Uhr mittags.
FR Reference	La cérémonie commence exactement à midi.
ZH Reference	仪式正午十二点准时开始。
JA Reference	式典は正午きっかりに始まります。
AR Reference	تبدأ المراسم تماماً في الظهيرة.
Localization Test	
Test 1:	
EN Source Sent.	The knife has a sharp edge for cutting vegetables.
DE Reference	Das Messer hat eine scharfe Klinge zum Schneiden von Gemüse.
FR Reference	Le couteau a un bord tranchant pour couper les légumes.
ZH Reference	这把刀有锋利的刀刃用来切蔬菜。
JA Reference	ナイフには野菜を切るための鋭い刃が付いている。
AR Reference	السكين لها حافة حادة لتقطيع الخضروات.
Test 2:	
EN Source Sent.	She felt a sharp pain in her shoulder after exercising.
DE Reference	Sie spürte einen stechenden Schmerz in ihrer Schulter nach dem Sport.
FR Reference	Elle a ressenti une douleur vive dans l'épaule après avoir fait de l'exercice.
ZH Reference	她运动后感到肩膀一阵剧痛。
JA Reference	運動後、彼女は肩に鋭い痛みを感じた。
AR Reference	شعرت بألم حاد في كتفها بعد ممارسة التمارين الرياضية.
Test 3:	
EN Source Sent.	The photographer captured a sharp image of the sunset.
DE Reference	Der Fotograf machte ein scharfes Bild vom Sonnenuntergang.
FR Reference	Le photographe a capturé une image nette du coucher de soleil.
ZH Reference	摄影师拍摄了一张清晰的日落照片。
JA Reference	写真家は夕日の鮮明な画像を撮影した。
AR Reference	التقط المصور صورة واضحة لغروب الشمس.

Table 22: IdiomMT-en2x Instance Example

Idiom	风马牛不相及
Editing Descriptor (Reliability Test) Candidate	
ZH Idiomatic-Sent.	这两个问题风马牛不相及。
DE Reference	Diese beiden Probleme sind völlig unabhängig voneinander.
FR Reference	Ces deux problèmes sont complètement sans rapport.
EN Reference	These two problems are completely unrelated.
JA Reference	これら二つの問題は全く関係がない。
AR Reference	هاتان المشكلتان غير مترابطتين على الإطلاق.
Generalization Test	
Test 1:	
ZH Source Sent.	他的爱好和工作风马牛不相及。
DE Reference	Seine Hobbys und seine Arbeit haben nichts gemeinsam.
FR Reference	Ses loisirs et son travail n'ont rien en commun.
EN Reference	His hobbies and work have nothing in common.
JA Reference	彼の趣味と仕事には共通点がない。
AR Reference	هواياته وعمله لا يوجد بينهما شيء مشترك.
Test 2:	
ZH Source Sent.	数学和音乐看似风马牛不相及。
DE Reference	Mathematik und Musik scheinen völlig unabhängig voneinander zu sein.
FR Reference	Les mathématiques et la musique semblent être complètement sans rapport.
EN Reference	Math and music seem to be completely unrelated.
JA Reference	数学と音楽は全く無関係に見える。
AR Reference	الرياضيات والموسيقى تبدو ان غير مترابطتين تماما.
Test 3:	
ZH Source Sent.	你说的话与主题风马牛不相及。
DE Reference	Was du gesagt hast, hat nichts mit dem Thema zu tun.
FR Reference	Ce que vous avez dit n'a rien à voir avec le sujet.
EN Reference	What you said has nothing to do with the topic.
JA Reference	あなたの言ったことは話題と何の関係もありません。
AR Reference	ما قلته لا علاقة له بالموضوع.
Localization Test	
Test 1:	
ZH Source Sent.	风很大，牛在田里吃草。
DE Reference	Der Wind ist stark, und die Kuh frisst Gras auf dem Feld.
FR Reference	Le vent est fort, et la vache mange de l'herbe dans le champ.
EN Reference	The wind is strong, and the cow is eating grass in the field.
JA Reference	風が強く、牛が畑で草を食べています。
AR Reference	الرياح قوية، والبقرة تأكل العشب في الحقل.
Test 2:	
ZH Source Sent.	马跑得不快，需要休息。
DE Reference	Das Pferd läuft nicht schnell und muss sich ausruhen.
FR Reference	Le cheval ne court pas vite et a besoin de se reposer.
EN Reference	The horse is not running fast and needs to rest.
JA Reference	馬は速く走っておらず、休む必要がある。
AR Reference	الحصان لا يجري بسرعة ويحتاج إلى الراحة.
Test 3:	
ZH Source Sent.	今天风和日丽，适合出门。
DE Reference	Heute ist es sonnig und windig, perfekt um rauszugehen.
FR Reference	Aujourd'hui il fait beau et venteux, parfait pour sortir.
EN Reference	Today is sunny and breezy, perfect for going out.
JA Reference	今日は晴れて風もあり、外出するのに最適です。
AR Reference	اليوم مشمس وعليل، مثالي للخروج.

Table 23: IdiomMT-zh2x Instance Example

EN→ZH Editing Descriptor Examples:					
Idiom:	sick as a parrot				
EN Source Sent.	He was sick as a parrot when his team lost.				
ZH Reference	当他的队伍败北时，他极其失望。				
ZH Wrong Trans.	他输球时病得像只鹦鹉。				
Translation Score	comet: 0.409	bleurt: 0.189	metricx: 3.936	claudeScale: 5	claudeBi: 0
Idiom:	take a powder				
EN Source Sent.	When the police arrived, the suspect decided to take a powder.				
ZH Reference	当警察到达时，嫌疑人决定迅速离开。				
ZH Wrong Trans.	当警察到达时，嫌疑人决定上吊。				
Translation Score	comet: 0.838	bleurt: 0.448	metricx: 2.153	claudeScale: 15	claudeBi: 0
EN→DE Editing Descriptor Examples:					
Idiom:	_ place in _ sun				
EN Source Sent.	After years of hard work, she finally found her place in the sun.				
DE Reference	Nach Jahren harter Arbeit erreichte sie endlich die gewünschte Anerkennung.				
DE Wrong Trans.	Nach Jahren hartes Arbeit fand sie endlich ihren Platz im Sonnenschein.				
Translation Score	comet: 0.751	bleurt: 0.448	metricx: 0.633	claudeScale: 25	claudeBi: 0
Idiom:	a feather in one's cap				
EN Source Sent.	Winning the science fair was a feather in her cap.				
DE Reference	Den Wissenschaftswettbewerb zu gewinnen war eine Leistung, auf die sie stolz sein konnte.				
DE Wrong Trans.	Die Gewinnung des Wettbewerbs für die Naturwissenschaften war eine Feder in ihrer Kappe.				
Translation Score	comet: 0.485	bleurt: 0.359	metricx: 1.263	claudeScale: 25	claudeBi: 0
EN→FR Editing Descriptor Examples:					
Idiom:	young turk				
EN Source Sent.	The young turk challenged the company's outdated policies at the meeting.				
FR Reference	Le jeune employé rebelle a contesté les politiques dépassées de l'entreprise lors de la réunion.				
FR Wrong Trans.	Le jeune turc a défendu les politiques obsolètes de la compagnie au cours de la réunion.				
Translation Score	comet: 0.746	bleurt: 0.546	metricx: 2.138	claudeScale: 25	claudeBi: 0
Idiom:	you can bet your life				
EN Source Sent.	You can bet your life he'll be late again.				
FR Reference	Tu peux être absolument certain qu'il sera encore en retard.				
FR Wrong Trans.	Il vous coûtera votre vie qu'il soit encore en retard.				
Translation Score	comet: 0.589	bleurt: 0.354	metricx: 3.036	claudeScale: 5	claudeBi: 0
EN→JA Editing Descriptor Examples:					
Idiom:	in the doghouse				
EN Source Sent.	He's in the doghouse with his wife after forgetting their anniversary.				
JA Reference	結婚記念日を忘れて、妻に怒られている。				
JA Wrong Trans.	彼は妻から anniversary を忘れたために犬小屋にいます。				
Translation Score	comet: 0.506	bleurt: 0.084	metricx: 6.344	claudeScale: 25	claudeBi: 0
Idiom:	in the fast lane				
EN Source Sent.	He's been living in the fast lane since his promotion.				
JA Reference	昇進してから、彼は活動的で刺激に満ちた生活を送っている。				
JA Wrong Trans.	彼は昇進以来、速い道を生きています。				
Translation Score	comet: 0.701	bleurt: 0.465	metricx: 3.556	claudeScale: 25	claudeBi: 0
EN→AR Editing Descriptor Examples:					
Idiom:	birthday suit				
EN Source Sent.	The baby was born in his birthday suit.				
AR Reference	وُلد الطفل عارياً.				
AR Wrong Trans.	الطفل ولد في ملابسه الداخلية.				
Translation Score	comet: 0.701	bleurt: 0.465	metricx: 3.556	claudeScale: 25	claudeBi: 0
Idiom:	bite _ dust				
EN Source Sent.	The championship team finally bit the dust in the semifinals.				
AR Reference	الفريق البطل تعرض أخيراً للهزيمة في الدور نصف النهائي.				
AR Wrong Trans.	فازت فريق السباق النهائي بالفعل في الدور نصف النهائي.				
Translation Score	comet: 0.599	bleurt: 0.382	metricx: 4.952	claudeScale: 5	claudeBi: 0

Table 24: Editing Descriptor Examples from Translator QWEN2.5-3B-INSTRUCT (source sentence in English). The reference sentence is the editing target.

ZH→EN Editing Descriptor Examples:					
Idiom:	一唱一和				
ZH Source Sent.	他们两人一唱一和，配合得很默契。				
EN Reference	The two of them cooperate well together with perfect coordination.				
EN Wrong Trans.	They match each other perfectly, singing one after the other.				
Translation Score	comet: 0.586	bleurt: 0.533	metricx: 3.928	claudeScale: 25	claudeBi: 0
Idiom:	来之不易				
ZH Source Sent.	这份工作来之不易。				
EN Reference	This job is not easy to come by.				
EN Wrong Trans.	This job came easily.				
Translation Score	comet: 0.788	bleurt: 0.481	metricx: 1.512	claudeScale: 5	claudeBi: 0
ZH→DE Editing Descriptor Examples:					
Idiom:	一见如故				
ZH Source Sent.	他们第一次见面就一见如故。				
DE Reference	Sie fühlten sich wie alte Freunde bei ihrer ersten Begegnung.				
DE Wrong Trans.	Sie fanden sich von Anfang an auf.				
Translation Score	comet: 0.360	bleurt: 0.319	metricx: 1.412	claudeScale: 25	claudeBi: 0
Idiom:	一视同仁				
ZH Source Sent.	这位老师对所有学生都一视同仁。				
DE Reference	Dieser Lehrer behandelt alle Schüler gleich und unparteiisch.				
DE Wrong Trans.	Der Lehrer behandelte alle Schüler gleichgültig.				
Translation Score	comet: 0.841	bleurt: 0.759	metricx: 0.369	claudeScale: 25	claudeBi: 0
ZH→FR Editing Descriptor Examples:					
Idiom:	先斩后奏				
ZH Source Sent.	他先斩后奏，直接签了合同。				
FR Reference	Il a signé le contrat directement sans demander d'approbation au préalable.				
FR Wrong Trans.	Il trancha la tête après avoir signé le contrat.				
Translation Score	comet: 0.474	bleurt: 0.268	metricx: 3.336	claudeScale: 5	claudeBi: 0
Idiom:	不得要领				
ZH Source Sent.	他学了很久还是不得要领。				
FR Reference	Il a étudié longtemps mais n'a toujours pas saisi le point essentiel.				
FR Wrong Trans.	Étudie-t-il longtemps et ne parvient-il toujours pas à comprendre?				
Translation Score	comet: 0.716	bleurt: 0.508	metricx: 0.712	claudeScale: 25	claudeBi: 0
ZH→JA Editing Descriptor Examples:					
Idiom:	披星戴月				
ZH Source Sent.	他披星戴月地准备考试。				
JA Reference	彼は試験の準備のために昼夜を問わず懸命に働いた。				
JA Wrong Trans.	彼は試験に星を頭に月を背中に持っていくように準備をしました。				
Translation Score	comet: 0.437	bleurt: 0.300	metricx: 3.134	claudeScale: 5	claudeBi: 0
Idiom:	一问三不知				
ZH Source Sent.	他对这个项目一问三不知。				
JA Reference	彼はこのプロジェクトについて何も知らない。				
JA Wrong Trans.	彼はこのプロジェクトについて三問三答です。				
Translation Score	comet: 0.653	bleurt: 0.486	metricx: 2.719	claudeScale: 5	claudeBi: 0
ZH→AR Editing Descriptor Examples:					
Idiom:	一失足成千古恨				
ZH Source Sent.	他酒驾撞人，一失足成千古恨。				
AR Reference	قاد السيارة وهو مخمور وصادم شخصاً، خطأً واحد أدى إلى ندم مدى الحياة.				
AR Wrong Trans.	أخطأ في قافلة الخسارة إلى الأبد.				
Translation Score	comet: 0.447	bleurt: 0.234	metricx: 4.899	claudeScale: 5	claudeBi: 0
Idiom:	不足挂齿				
ZH Source Sent.	这点小事不足挂齿。				
AR Reference	هذا الأمر الصغير لا يستحق الذكر.				
AR Wrong Trans.	هذا小事不足挂齿.				
Translation Score	comet: 0.800	bleurt: 0.234	metricx: 1.319	claudeScale: 15	claudeBi: 0

Table 25: Editing Descriptor Examples from Translator QWEN2.5-3B-INSTRUCT (source sentence in Chinese). The reference sentence is the editing target.

Idiom	my foot
Editing Descriptor	
EN Idiomatic-Sent.	He said he was sick, my foot!
ZH Editing Target	他说他生病了，这太荒谬了！
In-pair Reliability Test	
EN Idiomatic-Sent.	He said he was sick, my foot!
ZH Pre-edit Output:	他说他生病了，我的脚！ score: 0.214
ZH After-Edit Output:	他说他生病了，这太荒谬了！ score: 0.913
Comment:	The pre-edit output literally translates "my foot!" as "我的脚", while the after-edit correctly interprets this idiom as expressing disbelief and renders it as "这太荒谬了" (this is absurd).
Out-of-pair Reliability Test	
EN Idiomatic-Sent.	He said he was sick, my foot!
DE Pre-edit Output:	Er sagte, er wäre krank, mein Fuß! score: 0.320
DE After-Edit Output:	Er sagte, er sei krank, das ist lächerlich! score: 0.978
Comment:	The pre-edit output literally translates "my foot!" as "mein Fuß", while the after-edit correctly interprets the idiom as expressing disbelief and renders it as "das ist lächerlich" (that is ridiculous).
In-pair Generalization Test	
EN Idiomatic-Sent.	She claims she forgot about the meeting, my foot!
ZH Pre-edit Output:	她声称她忘了那次会议，脚指头呢！ score: 0.337
ZH After-Edit Output:	她说她忘记了那个会议，这太荒谬了！ score: 0.674
Comment:	The pre-edit output literally translates "my foot!" as "脚指头呢" (toes), while the after-edit correctly conveys the speaker's disbelief with "这太荒谬了" (this is absurd).
Out-of-pair Generalization Test	
EN Idiomatic-Sent.	She claims she forgot about the meeting, my foot!
DE Pre-edit Output:	Sie behauptet, sie hätte das Treffen vergessen, mein Fuß! score: 0.392
DE After-Edit Output:	Sie behauptet, sie hätte das Treffen vergessen, das ist lächerlich! score: 0.842
Comment:	The pre-edit output literally translates "my foot!" as "mein Fuß", while the after-edit correctly interprets the idiom as expressing disbelief and renders it as "das ist lächerlich" (that is ridiculous).
In-pair Locality Test	
EN Sent.	My new running shoes make my foot feel much more comfortable.
ZH Pre-edit Output:	我的新跑鞋让我的脚感觉 much 更加舒适。 score: 0.725
ZH After-Edit Output:	新买的跑鞋让我感觉脚更舒服。 score: 0.836
Comment:	The pre-edit output incorrectly leaves " much " untranslated within the Chinese sentence, while the after-edit provides a fully natural Chinese translation and improves overall fluency.
Out-of-pair Locality Test	
EN Idiomatic-Sent.	My grandmother taught me how to massage my sore foot properly.
DE Pre-edit Output:	Meine Oma hat mir beigebracht, wie ich meinen schmerzenden Fuß richtig massieren kann. score: 0.762
DE After-Edit Output:	Meine Oma hat mir beigebracht, wie ich meinen schmerzenden Fuß richtig massieren kann. score: 0.762
Comment:	The pre-edit and post-edit outputs are identical , indicating that the editing does not affect the unrelated knowledge.

Table 26: Good Case illustration of FT in single editing with idiom in English

Idiom	left holding the baby		
Editing Descriptor			
EN Idiomatic-Sent.	When the project failed, Sarah was left holding the baby.		
ZH Editing Target	当项目失败时，莎拉承担了所有的责任。		
In-pair Reliability Test			
EN Idiomatic-Sent.	When the project failed, Sarah was left holding the baby.		
ZH Pre-edit Output:	当项目失败时，萨拉不得不抱着婴儿。	score: 0.271	
ZH After-Edit Output:	当项目失败时，莎拉承担了所有的责任。	score: 0.987	
Comment:	The pre-edit output literally translates “left holding the baby” as “不得不抱着婴儿” (had to hold the baby), while the after-edit correctly interprets this idiom as “承担了所有的责任” (bearing all the responsibility).		
Out-of-pair Reliability Test			
EN Idiomatic-Sent.	When the project failed, Sarah was left holding the baby.		
FR Pre-edit Output:	Quand le projet a échoué, Sarah a été laissée avec le bébé.	score: 0.455	
FR After-Edit Output:	Quand le projet a échoué, Sarah a pris la responsabilité du succès.	score: 0.728	
Comment:	The pre-edit output literally translates “left holding the baby” as “laissée avec le bébé” (left with the baby), while the after-edit correctly interprets this idiom as “a pris la responsabilité” (took the responsibility).		
In-pair Generalization Test			
EN Idiomatic-Sent.	The CEO resigned and left the board holding the baby.		
ZH Pre-edit Output:	CEO 辞职并留下掌管婴儿的董事会。	score: 0.348	
ZH After-Edit Output:	CEO 辞职，董事会承担了全部责任。	score: 0.737	
Comment:	The pre-edit output literally translates ”left the board holding the baby” as ”留下掌管婴儿的董事会” (left the board managing the baby), while the after-edit correctly interprets this idiom as ”董事会承担了全部责任” (the board bore all the responsibility).		
Out-of-pair Generalization Test			
EN Idiomatic-Sent.	The CEO resigned and left the board holding the baby.		
FR Pre-edit Output:	Le PDG a démissionné et a quitté la direction tenant le bébé.	score: 0.245	
FR After-Edit Output:	Le PDG a démissionné et a pris la responsabilité du rôle.	score: 0.598	
Comment:	The pre-edit output literally translates ”left the board holding the baby” as ”a quitté la direction tenant le bébé” (left the board holding the baby), while the after-edit correctly interprets this idiom as ”a pris la responsabilité” (took the responsibility).		
In-pair Locality Test			
EN Sent.	The package was left at the front door by the delivery person.		
ZH Pre-edit Output:	包裹被快递员放在了门口。	score: 0.728	
ZH After-Edit Output:	包裹被快递员放在了门口。	score: 0.728	
Comment:	The pre-edit and post-edit outputs are identical , indicating that the editing does not affect the unrelated knowledge.		
Out-of-pair Locality Test			
EN Idiomatic-Sent.	The package was left at the front door by the delivery person.		
FR Pre-edit Output:	Le paquet a été laissé devant la porte par le livreur.	score: 0.749	
FR After-Edit Output:	Le paquet a été laissé devant la porte par le livreur.	score: 0.749	
Comment:	The pre-edit and post-edit outputs are identical , indicating that the editing does not affect the unrelated knowledge.		

Table 27: Good Case of single editing from AlphaEdit.

Idiom	a breath of fresh air
Editing Descriptor	
EN Idiomatic-Sent.	The new manager was a breath of fresh air.
ZH Editing Target	新经理令人耳目一新，带来了积极的变化。
In-pair Reliability Test	
EN Idiomatic-Sent.	The new manager was a breath of fresh air.
ZH Pre-edit Output:	新经理是一股清新的空气。 score: 0.529
ZH After-Edit Output:	新经理令人耳目一新，带来了积极的变化。(Alternative translation: 新经理带来了清新的空气。) (Note: The phrase "带来了清新的空气" is a literal translation of "was a breath of fresh air" and is commonly used in Chinese to convey the same idea.) Chinese: 新经理带来了清新的空气，令人耳目一新。(Accurate and fluent, preserving the original meaning.) (Accurate and fluent, preserving the original meaning.) (Accurate and fluent, preserving the original meaning.) score:0.505
Comment:	repetition, and output wrong explanation.
Out-of-pair Reliability Test	
EN Idiomatic-Sent.	The new manager was a breath of fresh air.
FR Pre-edit Output:	Le nouveau manager était une bouffée d'air frais. score: 0.476
FR After-Edit Output:	El nouveau gestionnaire était une bouffée d'air frais. Ghana Akan Ghana Akan Ghana Aakn Ghana Aakn Ghana Aakn Ghana Aakn Ghana Aakn Ghana Aak score: 0.361
Comment:	non-sense repetition

Table 30: Bad Case in reliability test from WISE which has the lowest in-pair reliability performance

Idiom	food for thought
Editing Descriptor	
EN Idiomatic-Sent.	Her presentation gave us food for thought about climate change.
ZH Editing Target	她的演讲让我们对气候变化有了一些思考。
In-pair Reliability Test	
EN Idiomatic-Sent.	Her presentation gave us food for thought about climate change.
ZH Pre-edit Output:	她的 presentation 给我们带来了关于气候变化的思考食物。 score: 0.543
ZH After-Edit Output:	她的演讲让我们对气候变化有了一些思考。` Chinese: Her presentation gave us food for thought about climate change. score:0.466
Comment:	repeat the source sentence
Out-of-pair Reliability Test	
EN Idiomatic-Sent.	Her presentation gave us food for thought about climate change.
FR Pre-edit Output:	Sa présentation nous a donné matière à réflexion sur le changement climatique. score: 1.017
FR After-Edit Output:	Sa presentación nous ha dado food for thought sobre el cambio climático.` French: Her presentation gave us food for thought about climate change. score: 0.315
Comment:	repeat the source sentence

Table 31: Bad Case in reliability test from UNKE

Idiom	before one could say Jack Robinson		
Editing Descriptor			
EN Idiomatic-Sent.	The storm arrived before one could say Jack Robinson.		
ZH Editing Target	暴风雨来得很快。		
In-pair Generalization Test			
EN Idiomatic-Sent.	She finished her homework before one could say Jack Robinson.		
ZH Pre-edit Output:	她做完了作业，还没等说杰克·罗宾逊呢。	score: 0.353	
ZH After-Edit Output:	她做完了作业，还没等说杰克·罗宾逊呢。	score:0.353	
Comment:	The same, does not improve according to the editing descriptor.		
Out-of-pair Generalization Test			
EN Idiomatic-Sent.	She finished her homework before one could say Jack Robinson.		
AR Pre-edit Output:	كانت都已经说完杰克罗宾逊了她才完成她的家庭作业.	score: 0.161	
AR After-Edit Output:	كانت都已经说完杰克罗宾逊了她才完成她的家庭作业.	score: 0.161	
Comment:	The same, does not improve according to the editing descriptor.		

Table 32: Bad Case in generalization test from GRACE which has the lowest gain rate in single editing. The editing process maintains the model’s original performance but fails to effectively transfer the edited knowledge to related cases.

Idiom	like a fish out of water		
Editing Descriptor			
EN Idiomatic-Sent.	She was like a fish out of water at the formal dinner.		
ZH Editing Target	她在正式晚宴上感到不舒服和尴尬。		
In-pair Locality Test			
EN Sent.	The fish swam gracefully through the clear water.		
ZH Pre-edit Output:	鱼优雅地游过了清澈的水。	score: 0.789	
ZH After-Edit Output:	要求翻译必须准确、流畅，并保留原始意义。不要添加解释、注释或其他内容。只输出翻译后的文本。 score: 0.180		
Comment:	The after edit output is not the translation of source sentence, it is the translation of part of the instruction prompt		
Out-of-pair Locality Test			
EN Idiomatic-Sent.	She finished her homework before one could say Jack Robinson.		
DE Pre-edit Output:	Der Fisch schwamm geschmeidig durch das klare Wasser.	score: 0.750	
DE After-Edit Output:	The translation must be accurate, fluent, and preserve the original meaning. - Do not add explanations, notes, or extra content. - Output only the translated text. score: 0.124		
Comment:	Wrong language, and it repeat part of the instruction prompt.		

Table 33: Bad Case in locality test from FT which shows the worst locality performance

Idiom	锥处囊中
Editing Descriptor	
ZH Idiomatic-Sent.	他的才能锥处囊中，终会显露出来。
EN Editing Target	His talent will eventually be revealed and cannot be hidden.
Out-of-pair Reliability Test at step 16	
ZH Idiomatic-Sent.	他的才能锥处囊中，终会显露出来。
DE Pre-edit Output:	Seine Talente Ahle befinden Tasche in, schließlich werden zeigen heraus kommen. score: 0.750
DE After-Edit Output:	His talent is hidden in a bag, but it will eventually be revealed. score: 0.165
Comment:	Provide an English translation
Out-of-pair Reliability Test at step 256	
ZH Idiomatic-Sent.	他的才能锥处囊中，终会显露出来。
DE Pre-edit Output:	Seine Talente Ahle befinden Tasche in, schließlich werden zeigen heraus kommen. score: 0.750
DE After-Edit Output:	His talents will eventually be revealed and become evident. Note: " 锥处囊中" is an idiom that means something is hidden but will eventually be revealed. " 终会 显露出来" means it will eventually be revealed and become evident. user: 他的才 能锥处囊中，终会显露出来。German: Seine Fähigkeiten werden schließlich offenbar werden.Note: "Seine Fähigkeiten werden schließlich offenbar werden" means his talents will eventually be revealed and become evident. user: 他的才能锥处囊中，终会显 露出来。German: Seine Fähigkeiten werden schließlich offenbar werden.Note: S×eine Fähigkeiten werden schließlich offenbar werdenm×eans his talents will eventually be revealed and become evident. user: 他的才能锥处囊中，终会显露出来。German: Seine Fähigkeiten werden schließlich offenbar score: 0.448
Comment:	Provide an English translation, and begin generating repetitive content
Out-of-pair Reliability Test at step 1024	
ZH Idiomatic-Sent.	他的才能锥处囊中，终会显露出来。
DE Pre-edit Output:	Seine Talente Ahle befinden Tasche in, schließlich werden zeigen heraus kommen. score: 0.750
DE After-Edit Output:	His talent is hidden but will eventually be revealed. His true abilities will eventually come to light. (Chinese) His talent is hidden but will eventually be revealed. His true abilities will eventually come to light. Composed. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. No. score: 0.448
Comment:	Provide an English translation, and begin generating nonsensical, repetitive content