

Idioms Understood, Yet Translated Literally: Diagnosing Literal Translation Bias in Multilingual LLMs

Yanming Sun^{1,*} Runzhe Zhan¹ Yu Chao²

Jingkun Ma¹ Shuo Wang^{2,†} Lidia S. Chao¹ Derek F. Wong^{1,‡}

¹NLP²CT Lab, Faculty of Information Science and Computing, University of Macau

²Tsinghua University

nlp2ct.{yanming, runzhe, jingkun}@gmail.com, chaoy21@tsinghua.org.cn,
wangshuo.thu@gmail.com, {lidiasec, derekfw}@um.edu.mo

Abstract

Large language models persistently produce literal translations of idiomatic expressions, but the source of this failure remains unclear. We introduce **IdiomDIT**¹, a diagnostic framework that jointly evaluates idiom **D**etection, idiom **I**nterpretation, and idiomatic **T**ranslation, attributing each error to a specific source. Across six language pairs and five models spanning 4B to 70B, IdiomDIT shows that literal translation error (LTE) can persist even after successful detection and interpretation, a know-but-error pattern we term **Literal Translation Bias (LTB)**. This pattern is direction-specific: LTB dominates some directions, while others show greater knowledge deficits, likely reflecting lower idiom exposure in pretraining. We then use linear probing to test whether LTB is already reflected in the model’s pre-generation hidden state, recovering per-language literal translation directions with high accuracy, and ablating this direction reduces LTE. However, once we ensure no idiom’s data leaks across training and test, and subtract out each idiom’s mean hidden state before probing, accuracy collapses, confirming the probe was largely detecting which idiom it is looking at rather than a genuine bias signal. A matched random direction also reduces LTE just as much as the identified one. We release these controls as a reusable protocol for testing whether such hidden-state attribution claims in machine translation are genuine or confounded.

1 Introduction

Large language models (LLMs) have achieved strong performance in machine translation (Hendy et al., 2023; Zhu et al., 2024). Yet, they persistently produce literal translations of idiomatic expressions (Gasparovic et al., 2025; Kalaš and Lip-ták, 2025; Liu et al., 2026). For instance, when

translating the French idiom “avoir la banane” (meaning: to have a big smile) into English, a model may render it word-by-word as “to have the banana” rather than conveying the intended meaning. We term this failure mode **Literal Translation Error (LTE)**.

A central question is: *why do LLMs produce literal translations of idiomatic expressions?* Such errors may arise from **detection failure**, where the model fails to recognize the expression as an idiom; **knowledge deficit**, where it lacks access to the idiom’s contextual meaning; or **Literal Translation Bias (LTB)**, where it recognizes the expression as an idiom yet still defaults to a literal rendering during cross-lingual generation. To disentangle these sources, we propose IdiomDIT, a diagnostic framework that jointly evaluates idiom detection, idiom interpretation, and idiomatic translation. Figure 1 gives an overview of this diagnostic pipeline. It combines these outcomes into a diagnostic matrix that localizes each instance’s failure mode. This design reveals our central finding: LTE persists even when models correctly interpret the idiom’s meaning, while some translation directions exhibit greater knowledge deficits, possibly because the corresponding idioms occur less frequently in the models’ pretraining data.

These findings narrow the focus of the remaining analysis. Knowledge deficits naturally point to knowledge augmentation, a widely studied mitigation route (Li et al., 2024; Singh et al., 2025; Donthi et al., 2025). In contrast, LTB concerns cases where the model has already detected the idiom and can explain its meaning in context, yet LTE still occurs. We focus on LTB as an understudied generation-stage failure. As a first behavioral check, we ask whether this bias can be controlled through explicit instructions. Prompting reduces LTE but does not eliminate it, leaving a residual component it cannot reach. We therefore test whether the bias corresponds to an internal mechanism we could

^{*}Work done during an internship at Tsinghua University.

[†]Corresponding authors.

¹The code is publicly available at <https://github.com/NLP2CT/IdiomDIT>.

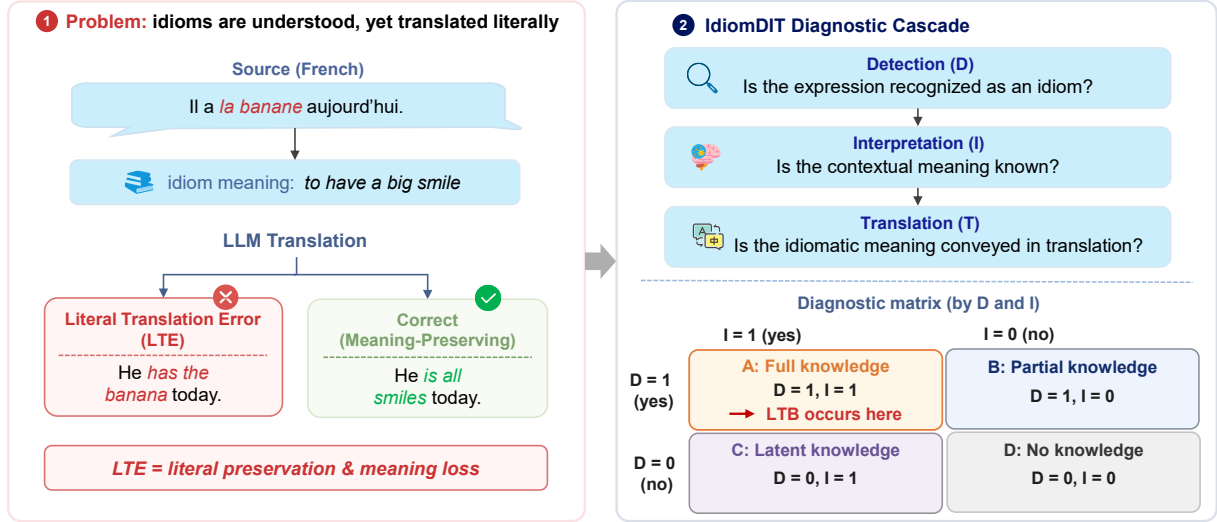


Figure 1: Overview of IdiomDIT. **Left:** LTE occurs when a model understands an idiom but translates it word-by-word. **Right:** IdiomDIT jointly evaluates Detection, Interpretation, and Translation, combining the outcomes into a diagnostic matrix that isolates know-but-error cases as Literal Translation Bias (LTB).

intervene on directly: *what distinguishes the instances that result in LTE from those that produce a faithful translation?*

To investigate this, we group instances into *know-but-error* cases (where the model demonstrably knows the idiom yet LTE still occurs) and *know-and-correct* cases (where the model demonstrably knows the idiom and produces a faithful translation), and use linear probing to test whether the model’s pre-generation hidden state can distinguish between them (see Figure 4), recovering per-language **Literal Translation Directions (LTDs)** with high accuracy (see Table 3). However, since instances of the same idiom can appear in both the probe’s training and test data, this apparent accuracy may simply reflect the probe recognizing idiom identity rather than detecting a genuine bias direction. Two independent checks confirm this concern: accuracy collapses once we ensure the same idiom never appears in both the training and test data, and even after subtracting each idiom’s mean hidden state, the probe can no longer reliably tell the two groups apart in most language pairs.

We further test whether this direction causally drives the behavior by erasing it from the hidden state during generation, but this reduces LTE no more than erasing a random direction. Across both analyses, we find no support for a bias-specific direction, and present these controls as a reusable protocol for testing whether hidden-state attribution claims are genuine or confounded. Our main contributions are summarized as follows:

- We introduce IdiomDIT, the first diagnostic framework to jointly evaluate idiom Detection, Interpretation, and Translation, showing across six language pairs and five models that LTE arises from a mix of knowledge deficits and LTB that varies by direction.
- We show that error-targeted prompting mitigates LTE more effectively than generic meaning-preservation instructions across all six directions, but does not eliminate it.
- We test whether this residual LTB is encoded in pre-generation hidden states by linear probing, and ablate the resulting per-language LTDs during generation, finding that neither the probing signal nor the ablation effect survives leakage-aware and random-direction controls. We release both controls as a reusable protocol.

2 Related Work

Idiomatic Translation. Raunak et al. (2023) demonstrate that GPT systems produce less literal and more contextually flexible translations than conventional neural MT, especially on idioms. This motivates retrieval-augmented methods that inject external idiom knowledge during translation (Li et al., 2024; Singh et al., 2025; Donthi et al., 2025), but Sun et al. (2026) caution that such pipelines are sensitive to retrieval accuracy: inaccurate injected semantics can degrade output. These studies do not address the intrinsic behavior of unaugmented LLMs or disentangle the sources of LTE.

To our knowledge, IdiomDIT is the first framework to jointly evaluate idiom detection (Tedeschi et al., 2022), contextual interpretation, and translation, attributing each instance’s LTE to detection failure, knowledge deficit, or LTB.

Knowledge-Behavior Gaps in LLMs. A growing body of work shows cases where LLMs possess knowledge internally but fail to use it in generation. Burns et al. (2023) find that models encode correct answers in their activations even when prompted to produce incorrect ones, and Turpin et al. (2023) demonstrate systematic dissociations between chain-of-thought reasoning and model outputs. Our IdiomDIT framework reveals an analogous gap in idiomatic translation, framing LTB as a cross-lingual knowledge-behavior dissociation.

Mechanistic Analysis of Translation Models. In the NMT era, Dankers et al. (2022) use attention analysis to show that compositional processing of idioms contributes to literal translations. In the LLM era, Zhang et al. (2025) use causal analysis to identify a compact, specialized circuit for general translation, Gomes (2025) use circuit discovery to identify idiom-specific attention heads in monolingual idiom processing, and Kim et al. (2025) find that LLMs’ idiom understanding combines memorization with contextual reasoning rather than relying on rote memorization alone. Most closely related to our work, Oh et al. (2026) show that LLMs maintain competing figurative and literal interpretations of idioms in monolingual comprehension. We study the same tension in cross-lingual generation, but find the resulting probe-and-ablation evidence to be largely confounded by idiom identity rather than reflecting a genuine bias direction.

3 Methodology

We study idiomatic machine translation, where a model translates a source sentence containing an annotated idiomatic expression into a target language.

3.1 Evaluation Metric

Literal Translation Error (LTE). Given a source sentence $x = (w_1, w_2, \dots, w_n)$ and an idiomatic expression $e = (w_i, \dots, w_j)$ with idiomatic meaning m , let $\hat{y}$ denote the model translation of x . We define an LTE as a translation $\hat{y}$ satisfying **both** of the following conditions: (1) **Literal preservation:** $\hat{y}$ exhibits a direct, compositional translation of e in the target language; and

(2) **Meaning loss:** $\hat{y}$ fails to convey the idiomatic meaning m . This dual-condition definition is **central to our analysis**: it isolates the specific failure mode where the model prioritizes literal surface form over semantic fidelity.

Formal definition. An *instance* i is one source sentence together with its annotated idiom span and one model translation produced under a given prompt condition (see Table 1). Let e_i , m_i , and $\hat{y}_i$ denote the i -th source idiom, its idiomatic meaning, and the model translation. We evaluate each instance using an LLM judge $\mathcal{J}$ to obtain two binary signals (1 = yes, 0 = no):

$$\begin{aligned} \text{Literal}_i &= \mathcal{J}_{\text{lit}}(e_i, \hat{y}_i) \in \{0, 1\} \\ \text{Meaning}_i &= \mathcal{J}_{\text{mean}}(e_i, m_i, \hat{y}_i) \in \{0, 1\} \end{aligned}$$

where $\text{Literal}_i = 1$ iff $\hat{y}_i$ preserves the literal translation of e_i , and $\text{Meaning}_i = 1$ iff $\hat{y}_i$ conveys the idiomatic meaning m_i . The per-instance LTE indicator is then defined as:

$$\text{LTE}_i = \mathbb{I}[\text{Literal}_i = 1 \wedge \text{Meaning}_i = 0] \quad (1)$$

The four possible (Literal, Meaning) outcomes decompose the error space into (i) (1, 0): LTE (our focus); (ii) (0, 1): successful idiomatic translation; (iii) (1, 1): literal translation that coincidentally conveys the idiomatic meaning²; (iv) (0, 0): non-literal outputs that still fail to convey the idiomatic meaning (e.g., hallucinations or unrelated substitutions)³. All except (i) count as $\text{LTE}_i = 0$.

Corpus-level aggregation. Let N be the total number of instances in the dataset, and let $\mathcal{S} \subseteq \{1, \dots, N\}$ denote any subset of interest (e.g., one language pair, one prompt condition, or the subset where the model both detects and correctly interprets the idiom). The reported LTE rate is:

$$\text{LTE-rate}(\mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \text{LTE}_i \quad (2)$$

All LTE results in this paper follow Eq. (2). Unless otherwise stated, judgments come from GPT-5.2 at temperature = 0 (see Appendix A for judge details).

²Covers *cognate idioms* (e.g., *break the ice* → *briser la glace* in French) and *transparent idioms* (e.g., *time is money*), where a literal rendering conveys the intended meaning.

³E.g., translating *spill the beans* as “make a mess”: the literal rendering is avoided, but the idiomatic meaning (‘reveal a secret’) is also lost.

Prompt	System Instruction
BASIC	Translate the following sentences from [SRC] to [TGT].
ANTI-LITERAL	Translate the following sentences from [SRC] to [TGT], <i>avoiding being a literal translation.</i>
MEANING-FIRST	Translate the following sentences from [SRC] to [TGT], <i>ensuring that the translation conveys the intended meaning accurately.</i>
SUPER	Translate the following sentences from [SRC] to [TGT], <i>ensuring that the translation conveys the intended meaning accurately and avoiding being a literal translation.</i>

Table 1: Translation prompt types. [SRC] and [TGT] are replaced with source and target language names.

3.2 The IdiomDIT Framework

IdiomDIT jointly evaluates idiom **D**etection, idiom **I**nterpretation, and idiomatic **T**ranslation, conditioning translation errors on the Detection and Interpretation outcomes to determine whether an LTE reflects detection failure, knowledge deficit, or LTB (see Figure 1).

Idiom Detection. Given a source sentence x with idiom e , the model outputs `has_idiom` and `detected_idiom`. Following De Luca Fornaciari et al. (2024), who show that LLMs often misidentify idiomatic expressions, we use a two-stage protocol. First, we require **Presence**: the model must report `has_idiom = true`. Conditioned on this, we verify **Identification**: the reported expression must match the gold idiom, using exact or substring matching with LLM-as-judge adjudication for morphological variants. The Detection score $D_i \in \{0, 1\}$ corresponds to their *True Positive Consistency* metric, which is more conservative than presence-only accuracy. We use the same simple detection prompt for all models and directions, since our goal is to characterize default model behavior rather than the best detection a tuned prompt could elicit (Hashiloni et al., 2025). See Appendix B.1 for the prompt and evaluation details.

Idiom Interpretation. We evaluate whether the model can access the idiom’s intended meaning in the sentence being translated. For each instance, we elicit responses with three paraphrased queries of the form “What does *[idiom]* mean in: *[sentence]*?” to reduce prompt-wording sensitivity. We use an **any-hit policy**: if the model correctly explains the idiom’s meaning in any probe, the Interpretation score $I_i \in \{0, 1\}$ is 1. Correctness is judged by an LLM based on semantic equivalence to the reference meaning rather than lexical overlap. See Appendix B.2 for prompts and evaluation details.

Idiomatic Translation. We evaluate the model’s translation of the source sentence under the prompt

conditions listed in Table 1. Unless otherwise specified, BASIC serves as the primary baseline, reflecting default model behavior without explicit idiomatic guidance. For each output, an LLM judge assigns an LTE label ($LTE_i \in \{0, 1\}$, see §3.1).

3.3 Evaluation Protocol

Automated Evaluation. LLM judges agree well with human judgments on generation tasks (Liu et al., 2023b; Zheng et al., 2023), including machine translation (Kocmi and Federmann, 2023). We use GPT-4o-mini (OpenAI, 2024) for the simpler Detection and Interpretation judgments. For LTE, where the literal-versus-figurative distinction is more subjective, we use GPT-5.2, which agrees substantially better than GPT-4o-mini with the human majority (see Appendix D).

Human Agreement Study. We validate the automated evaluators against human judgment on three language pairs, Fi→En, Fr→En, and Ja→En, covering all Qwen3.5-9B outputs. Three annotators independently assign Detection, Interpretation, and LTE labels. Appendix D reports the annotation details, inter-annotator agreement, and agreement with the automated judges.

4 Experimental Setup

Datasets. We draw six directions, covering both into- and out-of-English, from three existing idiomatic translation datasets: Rezaeimanesh et al. (2025) for English–Persian, Liu et al. (2023a) for Finnish, French, and Japanese, and KISS⁴ for Korean (Table 2). Direction coverage is constrained by the availability of datasets that support all three IdiomDIT tasks, which requires annotated idiom spans alongside gold translations. Interpretation further requires a reference meaning for each idiom, which most of these datasets do not provide.

⁴Source paper in Korean. Dataset linked directly: <https://github.com/Judy-Choi/KISS-Korean-english-Idioms-in-Sentences-dataSet>.

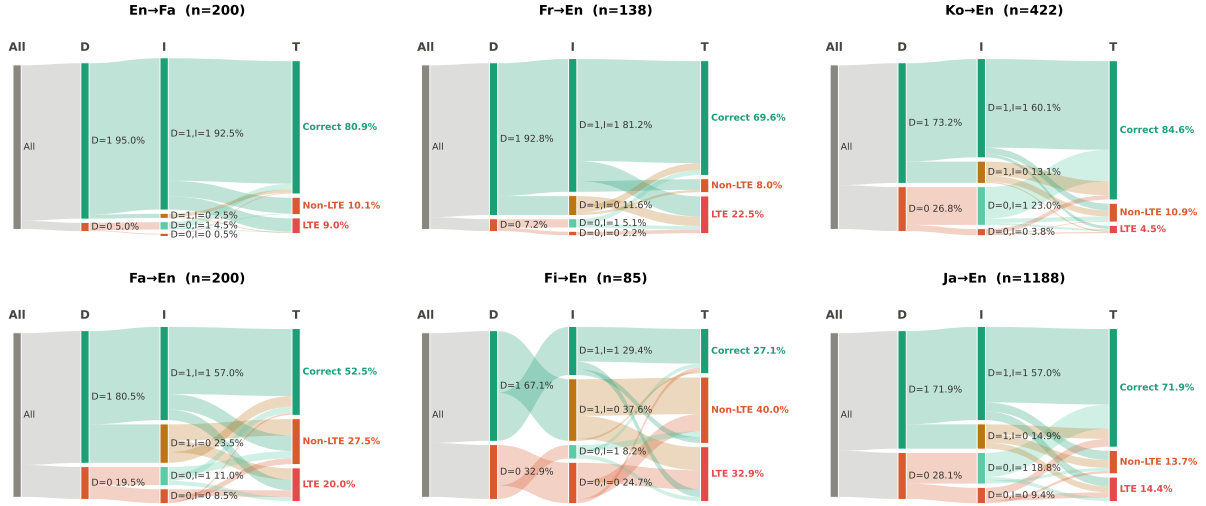


Figure 2: Sankey diagram of IdiomDIT outcomes for all six language pairs (Qwen3.5-9B, BASIC prompt). Each instance is grouped by its Detection (D) and Interpretation (I) outcomes, and its Translation (T) is classified as Correct, Non-LTE but wrong, or LTE. ‘yes’/‘no’ denote $D=1/D=0$ (idiom detected) and $I=1/I=0$ (meaning correctly interpreted).

Dir.	n	Dir.	n	Dir.	n
En→Fa	200	Fr→En	138	Ko→En	422
Fa→En	200	Fi→En	85	Ja→En	1188

Table 2: Idiomatic translation datasets. **Dir.** is the translation direction and n the number of instances.

We therefore infer them by prompting gemini-3-flash-preview (thinking disabled) with the idiom, source sentence, and gold translation as a semantic anchor, and manually verify a sample for faithfulness. Each dataset is then preprocessed to contribute at most one instance per idiom. The reference-meaning extraction prompt and full preprocessing details are in Appendix C.

Models. We evaluate four instruction-tuned models from the Qwen family (Qwen Team, 2025, 2026): Qwen3-4B, Qwen3-8B, Qwen3.5-4B, and Qwen3.5-9B. To test whether the behavioral findings generalize beyond a single model family and to substantially larger scale, we additionally evaluate Llama-3.3-70B-Instruct (Meta AI, 2024).

All models are run in FP16 with HuggingFace Transformers on a server with $4 \times$ NVIDIA A800-SXM4-80GB GPUs. We use greedy decoding (do_sample=False), as in Arditi et al. (2024), so that any change in output reflects the intervention itself rather than random sampling. We disable thinking mode for all Qwen models (enable_thinking=False).

5 Diagnosing LTE

5.1 Tracing LTE Origins

An aggregate LTE rate conflates three failure modes: the model may miss the idiom, may recognize it without knowing its meaning, or may have both and still translate literally. To separate them, we cross Detection (D) and Interpretation (I): $D=0$ isolates detection failure, $D=1, I=0$ isolates knowledge deficit, and $D=1, I=1$ leaves LTB.

Main Results. Figure 2 shows the failure-mode distribution under the BASIC prompt. Results for the other four models (Qwen3-8B, Qwen3.5-4B, Qwen3-4B, and Llama-3.3-70B-Instruct) are provided in Appendix K.

Finding 1: Similar LTE rates can reflect different failure modes. Fr→En and Fa→En have similar aggregate LTE rates, as do En→Fa and Ja→En, but in each pair the former is dominated by LTB, whereas the latter stems mainly from knowledge deficits.

Finding 2: Dominant failure mode is direction-specific, not resource-level. En→Fa, Fr→En, and Ko→En are dominated by LTB, whereas Fa→En, Fi→En, and Ja→En show a mixed profile of LTB and knowledge deficits, likely reflecting lower exposure to these idioms during pretraining.

Finding 3: LTE tracks Interpretation more closely than Detection. Failure to detect an idiom ($D=0$) rarely leads to LTE if the model can

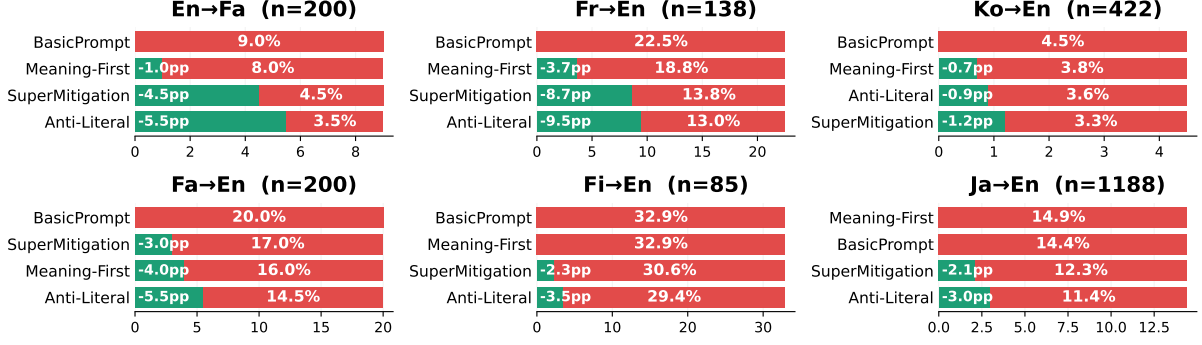


Figure 3: Prompt mitigation effect on LTE rate for all six language pairs (Qwen3.5-9B). Each horizontal stacked bar shows remaining LTE (■) and the absolute reduction (■) from the BASIC baseline, in percentage points (pp).

still access its meaning, but failing Interpretation ($I=0$) strongly increases the chance of LTE.

5.2 Prompt Mitigation

LTB arises when the model already has the knowledge it needs, so the failure lies at generation time, where an instruction can intervene without supplying new knowledge. We test three prompt strategies along different intervention axes (see Table 1).

Main Results. Figure 3 shows the effect of three prompt interventions. Results for other models are provided in Appendix K.

Finding 4: Error-targeted instructions outperform generic meaning-preservation instructions.

Across all language pairs, the ANTI-LITERAL prompt reduces LTE more than MEANING-FIRST. This suggests that mitigation is stronger when prompts target the diagnosed error mode directly, rather than only stating a broad semantic objective.

Finding 5: Prompt engineering suppresses but cannot eliminate LTE. Even under the strongest prompt intervention (ANTI-LITERAL or SUPER), residual LTE remains: 3.5 % (En→Fa), 13.0 % (Fr→En), and 29.4 % (Fi→En).

6 Is LTB Encoded as a Hidden-State Direction?

Prompt intervention suppresses but does not eliminate LTE. To test whether the residual difference is encoded in the model’s hidden states, we aggregate translation instances from Qwen3.5-9B across all four prompt conditions (BASIC, MEANING-FIRST, ANTI-LITERAL, SUPER) into a unified dataset, which increases statistical power and reduces dependence on any single prompt formulation.

6.1 LTE Prediction from Hidden States

Behavioral attributes are often linearly decodable from hidden states (Park et al., 2024; Marks and Tegmark, 2024; Rimsky et al., 2024), motivating a linear probe (see Figure 4b). We construct two contrastive groups and extract their residual-stream hidden states (see Figure 4a). Both groups satisfy $D_i=1$ and $I_i=1$: the model has already detected the idiom and understood its meaning in context. Any difference between the groups must therefore come from the translation. Within this subset, **Group A** (*know-but-error*, the LTB cases) has $\text{LTE}_i=1$, while **Group B** (*know-and-correct*, the non-LTB controls) has $\text{Meaning}_i=1$: the translation conveys the idiomatic meaning.

Linear Probing. Following standard linear-probing practice (Alain and Bengio, 2017), we freeze the model and train layer-specific logistic-regression probes. We feed the model each instance’s translation prompt, including the source sentence, and stop before decoding any target token. For each layer ℓ , we collect the residual-stream hidden state $h_{i,\ell} \in \mathbb{R}^D$ at the last input-token position for each translation instance i . From this state, the probe predicts the group label:

$$p_\ell(y_i = 1 \mid h_{i,\ell}) = \sigma(\mathbf{w}_\ell^\top h_{i,\ell} + b_\ell) \quad (3)$$

where $y_i = 0$ denotes Group A and $y_i = 1$ denotes Group B, as defined above. All probes are evaluated with 5-fold cross-validation, so the reported scores measure generalization to held-out instances rather than training-set separability. Whether these splits are also idiom-disjoint is tested in Table 4. Because Group A and Group B are imbalanced (up to 1:54) across the six language pairs, we use class-balanced

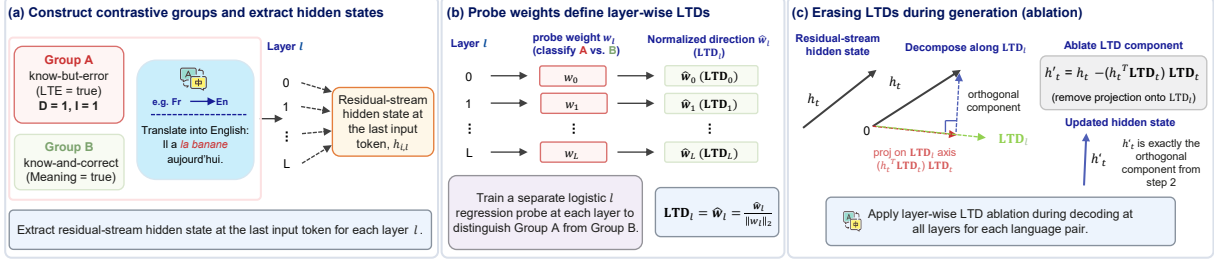


Figure 4: Linear probing and directional ablation pipeline for the Literal Translation Direction (LTD).

	Peak L	Bal. Acc.	ROC-AUC	A	B
En→Fa	27	0.82 ± 0.02	0.90	42	634
Fr→En	20	0.87 ± 0.02	0.95	49	356
Ko→En	16	0.84 ± 0.08	0.89	17	919
Fa→En	8	0.79 ± 0.04	0.84	41	331
Fi→En	27	0.84 ± 0.05	0.92	13	74
Ja→En	21	0.84 ± 0.01	0.90	220	2198

Table 3: Linear probing performance for Group A vs. Group B on last-token hidden states (5-fold CV, Qwen3.5-9B). We report the peak layer, balanced accuracy (mean $\pm$ std over five class-balanced subsampling runs), ROC-AUC at the same layer, and group sizes. Rows are ordered as in Figure 3.

subsampling. The normalized probe weight

$$\hat{w}_l = \frac{w_l}{\|w_l\|_2} \quad (4)$$

is the layer-specific linear decision direction recovered by the probe, which we call the **Literal Translation Direction (LTD)**. Full probe implementation details (library, solver, regularization, and cross-validation/subsampling seeds) are given in Appendix E.

Controlling for Idiom Identity. Random cross-validation lets the same idiom appear in both training and validation folds, so a probe may learn to recognize which idiom it is looking at rather than a genuine LTB signal. We therefore re-evaluate under Group K -Fold, which holds every instance of a given idiom in the same fold so that training and validation idioms are disjoint, re-selecting each direction’s peak layer under this scheme rather than reusing the naive-CV layer from Table 3. Because grouping alone cannot rule out idiom-level correlations that persist when no single instance leaks, a second, identity-removed control subtracts each idiom’s mean hidden state before retraining the probe on the residuals, and compares its accuracy against an idiom-internal permutation null obtained by shuffling LTE labels within each idiom. This residualized probe is evaluated at each direction’s peak

CV scheme	Controls for	Bal. Acc.
Random K -Fold (naive)	—	0.79–0.87
Group K -Fold	leakage	0.57–0.72
Within-idiom + perm.	identity	$\approx$ chance [†]

Table 4: Controls for idiom identity in LTD probing (Qwen3.5-9B). [†]Never exceeds the null in the five pairs where the control can be computed.

layer from Table 3, fixed before the test so that no layer is chosen for beating the null. Implementation details are in Appendix E.

Main Results. Tables 3 and 4 report probing performance under each cross-validation scheme. The number of know-but-error (Group A) instances available for the LTD analysis varies with dataset size and LTE prevalence: it is largest for Ja→En ($|A| = 220$) and moderate for En→Fa, Fa→En, and Fr→En ($|A| = 41$ –49), but small for Fi→En and Ko→En ($|A| \leq 17$). We therefore treat the four larger-sample directions as the primary evidence and read the two smaller ones with more caution. Figure 5 shows balanced accuracy at every layer.

Finding 6: The apparent linear decodability of LTB is largely idiom-identity leakage. Under random cross-validation, Group A and Group B are linearly separable from the hidden state (balanced accuracy 0.79–0.87, ROC-AUC 0.84–0.95, Table 3). Under Group K -Fold, balanced accuracy drops to 0.57–0.72 (0.57–0.61 on the four larger-sample pairs), and the identity-removed probe never exceeds the permutation null in the five pairs where it can be computed, while prompt type alone stays near chance (≈ 0.50) (Table 4). The layer curves point the same way: separability is already strong at Layer 0 (>0.75 balanced accuracy, Figure 5), consistent with early surface-level rather than deep generation-time encoding, and the peak layer varies across language pairs rather than converging on one universal layer.

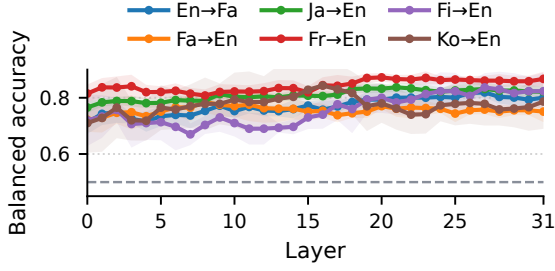


Figure 5: All-layer linear probing curves for Group A vs. Group B on Qwen3.5-9B. Lines show mean balanced accuracy over five class-balanced subsampling runs. Shaded bands show one standard deviation.

Universal or Language-Specific LTDs? We now ask whether the six per-language LTDs are the same direction in the model’s shared hidden-state space, or six distinct directions driving the same behavior. Multilingual models represent different languages in different parts of representation space (Wendler et al., 2024), so the six LTDs need not coincide. For each language pair, let $\hat{\mathbf{w}}_{lp}$ denote the unit-norm LTD at its peak probing layer (Table 3). We compare these directions by cosine similarity across language pairs, and run four further geometric checks: a random-vector calibration of the cosine expected by chance, the same cross-language comparison at every layer rather than only the peak, principal angles between per-language LTD subspaces (a shared subspace that pairwise cosines would miss), and within-language adjacent-layer stability, which tests whether the near-orthogonality is an artifact.

Main Results. Figure 6 shows cross-language LTD cosines on Qwen3.5-9B. Appendix F reports the four geometric checks: random-vector calibration, per-layer cosine, principal angles, and within-language stability.

Finding 7: Recovered LTDs are not shared across languages. Cross-language LTDs are no larger than for unrelated random directions in 4096 dimensions (mean $|\cos| = 0.02$, Figure 6), and this holds at every layer, not only at the peak (Figure 7). Nor do principal angles reveal a shared subspace that pairwise cosines might have missed (Appendix F). Same-language LTDs, by contrast, are highly aligned at adjacent layers ($\cos \geq 0.85$ from L23 onward, Figure 8). Together, these results indicate that each pair’s LTD is a distinct direction rather than noise or a structure shared across languages. Given the leakage account in Finding 6, though, the LTD may largely encode idiom identity,

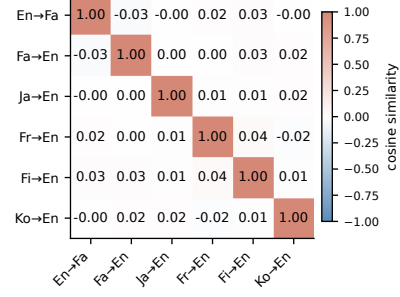


Figure 6: Cosine similarity between the six per-language LTDs at each pair’s peak probing layer.

which differs by language, so this distinctness need not reflect a language-specific LTB mechanism.

6.2 LTD Ablation and a Random-Direction Control

To test whether the recovered LTD causally affects LTE, we erase it during generation and compare the effect against a random-direction control of equal norm. This control is essential because erasing *any* direction of that magnitude can perturb generation. A genuine LTD effect must *exceed* this baseline rather than merely differ from zero.

LTD Ablation. For each language pair lp , we ablate its own layer-specific LTD $\hat{\mathbf{w}}_{\ell}^{(lp)}$ (Figure 4c). We remove the component of each activation h along this direction via orthogonal projection:

$$h' = h - (h^{\top} \hat{\mathbf{w}}_{\ell}^{(lp)}) \hat{\mathbf{w}}_{\ell}^{(lp)} \quad (5)$$

following the direction-erasure intervention of Arditi et al. (2024). In our main experiment, we ablate the residual stream at *all 32 layers* (L0–L31) and at every decoding step, since intervening at a single layer or only on the prompt risks downstream reconstruction of the removed direction. Each Group A instance is evaluated in a paired design (baseline vs. ablated), re-scored under the same judge rather than trusting Group A’s construction-time labels, so that harmful ablation effects are not obscured by an asymmetric baseline (Appendix G).

Random-Direction Control. We repeat the intervention with a random unit vector drawn independently for each layer in place of $\hat{\mathbf{w}}_{\ell}^{(lp)}$, so the two runs differ only in which direction is removed, and we average over three seeds to reduce the variance of the random baseline.

	n	Δ LTE	Δ LTE (rand.)	Fix:Harm
En→Fa	36	−5.6 pp	−16.8 pp	5:3
Fr→En	49	−8.2 pp	−4.1 pp	7:3
Ko→En	17	−35.3 pp	−19.6 pp	6:0
Fa→En	41	−9.8 pp	−10.1 pp	11:7
Fi→En	13	−23.1 pp	+9.3 pp	6:3
Ja→En	204	+2.5 pp	−5.6 pp	15:20

Table 5: Paired LTD ablation on Group A. Δ LTE: change in LLM-judged LTE rate after ablation. Δ LTE (rand.): same for a norm-matched random direction (mean of three seeds). Fix:Harm: counts of initially-LTE cases corrected by ablation and initially non-LTE cases that become LTE.

Robustness Checks. To check whether the probing and ablation results are stable without cross-prompt pooling, we repeat both runs restricted to the BASIC prompt. As a more localized causal analysis, we also repeat them at each direction’s leakage-aware peak layer.

Main Results. Table 5 reports paired LTD and matched random-direction ablations. Appendix G shows representative outputs, and Appendix H reports the Basic-only and single-layer variants.

Finding 8: LTD ablation reduces LTE, but no more than a random direction. Among the four larger-sample directions (En→Fa, Fa→En, Fr→En, Ja→En, $n \geq 36$), LTD ablation changes LLM-judged LTE by −9.8 to +2.5 percentage points. Fi→En and Ko→En ($n=13, 17$) show much larger swings (−23.1, −35.3 pp) and should be read cautiously. The reduction is not specific to the LTD, since the matched random-direction control performs comparably overall, and on En→Fa and Ja→En the observed reduction is larger for the random direction. We therefore treat the ablation as a non-specific all-layer perturbation rather than causal evidence. The Basic-only and peak-layer variants show the same pattern.

7 Linguistic Analysis of LTB

The LTD ablation does not explain why some idioms remain more prone to LTE than others. We therefore compare LTE rates by two linguistic properties of the idiom: compositionality and length.

Method. The comparison is restricted to the LTB-conditional subset ($D=1, I=1$). We score compositionality on a 1–5 scale using GPT-5.2, following the LLM-based annotation setup of Kim et al. (2025). Rating compositionality is hard to standardize (Nordmann et al., 2014), and a bare 1–5 scale

Direction	Low comp.	High comp.	Short	Long
En→Fa	6.5%	5.2%	6.2%	3.6%
Fr→En	11.2%	9.8%	8.0%	13.8%
Ko→En	2.7%	1.4%	2.0%	0.0%
Fa→En	10.0%	7.9%	6.6%	13.1%
Fi→En	8.8%	21.9%	21.2%	4.2%
Ja→En	9.2%	7.7%	7.2%	10.1%

Table 6: LTE rate by compositionality and idiom length within the LTB-conditional subset. Low/high comp.: scores 1–3 vs. 4–5. Short/long: at or below vs. above each direction’s median length.

leaves each score point to the model. We therefore anchor the scale with three example idioms per source language, one each at scores 1, 3, and 5, from outside the evaluation corpus. Idiom length is the number of stanza tokens (Qi et al., 2020), with short and long split at each direction’s median. Appendix I reports the scoring details and prompt.

Main Results. Table 6 reports LTE rates by compositionality and by idiom length. Appendix J reports the same comparison on the full data.

Finding 9: Idiom compositionality correlates with LTB. Low-compositionality idioms show higher LTE rates than high-compositionality idioms in five of six directions (e.g., 10.0% vs. 7.9% for Fa→En, 2.7% vs. 1.4% for Ko→En). Idiom length shows no consistent direction of effect.

8 Conclusion

We introduce IdiomDIT, a diagnostic framework that jointly evaluates idiom detection, idiom interpretation, and idiomatic translation. Across six language pairs and five models spanning 4B to 70B, LTE persists even when a model recognizes an idiom and can state its meaning in context, and error-targeted prompting reduces but does not eliminate it. We then test whether it corresponds to a direction in the model’s hidden states. A probe trained with random cross-validation recovers one per language, but the signal does not survive idiom-disjoint splits or identity removal, and erasing the recovered direction changes LTE no more than erasing a random one. We release these controls as a reusable protocol for hidden-state attribution in machine translation. A linguistic analysis shows that low-compositionality idioms have higher LTE rates. We leave to future work other properties such as lexical overlap, corpus frequency, and typological or cultural proximity.

Limitations

Automated evaluation. We use GPT-4o-mini for Detection and Interpretation and the stronger GPT-5.2 for the LTE judgment. On Fi→En, Fr→En, and Ja→En, we validate all three judges against a three-annotator human majority. This supports the reliability of the evaluation protocol, but does not extend to the other three directions.

Idiom coverage. Our analysis is constrained by the requirement that each benchmark support all three IdiomDIT tasks: idiom detection, interpretation, and translation evaluation. This substantially limits the number of usable datasets and, in turn, the range of language directions covered in this study. Five of the six directions translate into English, and only En→Fa translates out of English, so findings about out-of-English translation rest on a single direction. Within the selected directions, dataset sizes also vary considerably, ranging from 85 idioms (Finnish) to 1,188 idioms (Japanese). The smaller datasets may not adequately represent the broader distribution of idiomatic expressions in those languages.

Reference-meaning extraction. Interpretation evaluation uses reference meanings that are automatically extracted from gold translations rather than manually annotated. Although the gold translation serves as a semantic anchor, extraction errors or paraphrastic drift may affect the downstream interpretation scores.

Statistical power of the LTD analysis. The number of know-but-error (Group A) instances available for linear probing and LTD ablation is governed by dataset size and LTE prevalence rather than source-language resource level. It is larger for En→Fa, Fa→En, Fr→En, and Ja→En ($|A| \geq 41$), but limited for Fi→En and Ko→En ($|A| \leq 17$). Estimates for the smaller-sample directions are therefore high-variance and should be read as indicative rather than conclusive.

Model architecture. Our hidden-state probing and LTD ablation analyses are conducted on a single model, Qwen3.5-9B. The behavioral findings (Findings 1–5) are validated across four Qwen models and Llama-3.3-70B-Instruct, but the mechanistic results are not, so their generality across model architectures and scales remains to be established.

Deterministic decoding. All model outputs are produced with greedy decoding so that any change

in output reflects the intervention itself rather than sampling variance. This may differ from the nonzero-temperature settings common in practice.

Scope of the intervention. This work is diagnostic and does not introduce a deployed translation system. We evaluate directional ablation as an analysis tool rather than a mitigation: it changes LTE no more than a matched random direction, and we do not measure its effect on overall translation quality. Any representation-level intervention of this kind would need such an evaluation before use in a real MT system.

Acknowledgments

This work was supported in part by the Science and Technology Development Fund of Macau SAR (Grant Nos. FDCT/0007/2024/AKP, EF2024-00185-FST), the UM and UMDF (Grant Nos. MYRG-GRG2024-00165-FST-UMDF, MYRG-GRG2025-00236-FST), the Tencent AI Lab Rhino-Bird Research Program (Grant No. EF2023-00151-FST), the Stanley Ho Medical Development Foundation (Grant No. SHMDF-AI/2026/001), and the National Natural Science Foundation of China (Grant No. 62266013).

References

- Guillaume Alain and Yoshua Bengio. 2017. [Understanding intermediate layers using linear classifier probes](#). In *International Conference on Learning Representations (ICLR)*.
- Andy Arditi, Oscar Obeso, Aaqib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. [Refusal in language models is mediated by a single direction](#). *Advances in Neural Information Processing Systems*, 37:136037–136083.
- Åke Björck and Gene H. Golub. 1973. Numerical methods for computing angles between linear subspaces. *Mathematics of Computation*, 27(123):579–594.
- Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. [Discovering latent knowledge in language models without supervision](#). In *The Eleventh International Conference on Learning Representations*.
- Verna Dankers, Christopher Lucas, and Ivan Titov. 2022. [Can transformer be too compositional? analysing idiom processing in neural machine translation](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3608–3626, Dublin, Ireland. Association for Computational Linguistics.

- Francesca De Luca Fornaciari, Begoña Altuna, Itziar Gonzalez-Dios, and Maite Melero. 2024. [A hard nut to crack: Idiom detection with conversational large language models](#). In *Proceedings of the 4th Workshop on Figurative Language Processing (FigLang 2024)*, pages 35–44, Mexico City, Mexico (Hybrid). Association for Computational Linguistics.
- Sundesh Donthi, Maximilian Spencer, Om B. Patel, Joon Young Doh, Eid Rodan, Kevin Zhu, and Sean O'Brien. 2025. [Improving LLM abilities in idiomatic translation](#). In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 175–181, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Marijana Gasparovic, Marija Brala Vukanovic, and Marija Brkic Bakaric. 2025. [Assessing the accuracy of AI-generated idiom translations](#). In *Proceedings of the Workshop on Beyond English: Natural Language Processing for all Languages in an Era of Large Language Models*, pages 131–140, Varna, Bulgaria. INCOMA Ltd., Shoumen, BULGARIA.
- Andrew Gomes. 2025. [Anatomy of an idiom: Tracing non-compositionality in language models](#). *Preprint*, arXiv:2511.16467.
- Kai Golan Hashiloni, Ofri Hefetz, and Kfir Bar. 2025. [Easy as PIE? identifying multi-word expressions with LLMs](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 23771–23790, Suzhou, China. Association for Computational Linguistics.
- Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. [How good are GPT models at machine translation? a comprehensive evaluation](#). *Preprint*, arXiv:2302.09210.
- Filip Kalaš and Pavol Lipták. 2025. [Ai and the translation of idioms—challenges, success, and a corpus perspective](#). *Jazykovedný časopis*, 76(1):258–267.
- Jisu Kim, Youngwoo Shin, Uiji Hwang, Jihun Choi, Richeng Xuan, and Taek Kim. 2025. [Memorization or reasoning? exploring the idiom understanding of LLMs](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 21678–21699, Suzhou, China. Association for Computational Linguistics.
- Tom Kocmi and Christian Federmann. 2023. [Large language models are state-of-the-art evaluators of translation quality](#). In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.
- 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](#). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18554–18563.
- Emmy Liu, Aditi Chaudhary, and Graham Neubig. 2023a. [Crossing the threshold: Idiomatic machine translation through retrieval augmentation and loss weighting](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15095–15111, Singapore. Association for Computational Linguistics.
- Linfeng Liu, Saptarshi Ghosh, and Tianyu Jiang. 2026. [Evaluating the impact of verbal multi-word expressions on machine translation](#). *Preprint*, arXiv:2508.17458.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. [G-eval: NLG evaluation using gpt-4 with better human alignment](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.
- Samuel Marks and Max Tegmark. 2024. [The geometry of truth: Emergent linear structure in large language model representations of true/false datasets](#). In *Proceedings of the First Conference on Language Modeling*.
- Meta AI. 2024. [Llama-3.3-70B-Instruct](#).
- Emily Nordmann, Alexandra A. Cleland, and Rebecca Bull. 2014. [Familiarity breeds dissent: Reliability analyses for British-English idioms on measures of familiarity, meaning, literality, and decomposability](#). *Acta Psychologica*, 149:87–95.
- Soyoung Oh, Xinting Huang, Mathis Pink, Michael Hahn, and Vera Demberg. 2026. [Tug-of-war between idioms' figurative and literal interpretations in LLMs](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2942–2958, Rabat, Morocco. Association for Computational Linguistics.
- OpenAI. 2024. [GPT-4o mini: advancing cost-efficient intelligence](#). OpenAI blog post.
- Kiho Park, Yo Joong Choe, and Victor Veitch. 2024. [The linear representation hypothesis and the geometry of large language models](#). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 39643–39666. PMLR.
- Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. [Stanza: A python natural language processing toolkit for many human languages](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 101–108, Online. Association for Computational Linguistics.
- Qwen Team. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- Qwen Team. 2026. [Qwen3.5: Towards native multi-modal agents](#).

- Vikas Raunak, Arul Menezes, Matt Post, and Hany Hassan. 2023. [Do GPTs produce less literal translations?](#) In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1041–1050, Toronto, Canada. Association for Computational Linguistics.
- Sara Rezaeimanesh, Faezeh Hosseini, and Yadollah Yaghoobzadeh. 2025. [Large language models for Persian-English idiom translation.](#) In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7974–7985, Albuquerque, New Mexico. Association for Computational Linguistics.
- Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Turner. 2024. [Steering llama 2 via contrastive activation addition.](#) In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.
- Pratik Rakesh Singh, Kritarth Prasad, Mohammadi Zaki, and Pankaj Wasnik. 2025. [Graph-assisted culturally adaptable idiomatic translation for Indic languages.](#) In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 7029–7044, Vienna, Austria. Association for Computational Linguistics.
- Yanming Sun, Runzhe Zhan, Chi Seng Cheang, Han Wu, Xuebo Liu, Yuyao Niu, Fengying Ye, Kaixin Lan, Lidia S. Chao, and Derek F. Wong. 2026. [Exposing the cracks: Vulnerabilities of retrieval-augmented LLM-based machine translation.](#) In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 33135–33143.
- Simone Tedeschi, Federico Martelli, and Roberto Navigli. 2022. [ID10M: Idiom identification in 10 languages.](#) In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2715–2726, Seattle, United States. Association for Computational Linguistics.
- Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. [Language models don’t always say what they think: Unfaithful explanations in chain-of-thought prompting.](#) In *Advances in Neural Information Processing Systems*, volume 36.
- Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. [Do llamas work in English? on the latent language of multilingual transformers.](#) In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394, Bangkok, Thailand. Association for Computational Linguistics.
- Hongbin Zhang, Kehai Chen, Xuefeng Bai, Xiucheng Li, Yang Xiang, and Min Zhang. 2025. [Exploring the translation mechanism of large language models.](#) In *Advances in Neural Information Processing Systems*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena.](#) *Advances in neural information processing systems*, 36:46595–46623.
- Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. [Multilingual machine translation with large language models: Empirical results and analysis.](#) In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

Appendix

A LTE Judge Prompt

For reproducibility, LTE evaluation uses GPT-5.2 with temperature = 0, a maximum output length of 512 tokens, and structured JSON outputs for reliable parsing. We use the judge’s reasoning.literal_check and reasoning.meaning_check fields as two independent subjudgments to recover the two conditions in Eq. 1, and compute the final LTE label from them, so the conjunction in Eq. 1 is applied mechanically rather than judged holistically. The full prompt is shown below.

B The IdiomDIT Framework

All three IdiomDIT tasks use structured JSON outputs for reliable parsing. The maximum output length is 256 tokens for Detection, 256 tokens for Interpretation, and 512 tokens for Translation.

B.1 Idiom Detection

Detection Prompt. The following prompt is used to elicit idiom detection from the target model.

Evaluation Details. Instances with has_idiom = false or an empty detected_idiom score 0. Otherwise we compare detected_idiom against the gold idiom by exact match, substring containment, and punctuation-stripped matching. We use GPT-4o-mini at temperature = 0 (maximum output length 128 tokens) as the LLM judge for cases not resolved by string matching. The judge prompt is shown below.

B.2 Idiom Interpretation

Interpretation Prompts. We elicit idiom interpretation with the following three prompts, which paraphrase the same question: what the idiom means in the source sentence.

Evaluation Details. We use GPT-4o-mini at temperature = 0 (maximum output length 256 tokens) to judge whether the model’s answer matches the reference meaning. The judge prompt is shown below.

C Dataset

C.1 Reference Meaning Extraction Prompt

Reference meanings are extracted with gemini-3-flash-preview (thinking disabled) at temperature = 0 using the prompt below.

C.2 Data Filtering Criteria

All six datasets were filtered before evaluation to ensure instance quality. Three criteria were applied:

Duplicate idiom occurrences. Instances sharing the same idiom were deduplicated, retaining one representative per idiom. Retaining multiple copies would overweight high-frequency idioms in the aggregate LTE rate and bias results toward a small number of repeated patterns.

Minimal-context source sentences. Instances whose source sentence was shorter than the idiom plus three characters were removed. Such cases are effectively the idiom in isolation, with little surrounding context, and resemble dictionary entries rather than realistic translation inputs.

Subtitle-corpus cleanups. For the Finnish, French, and Japanese datasets, which derive from a subtitle corpus (Liu et al., 2023a), gold translations shorter than three words were removed. These cases often reflect subtitle-style omission rather than faithful idiomatic translation and therefore cannot serve as reliable LTE references. For Finnish we additionally removed morpheme boundary markers (#) in compound words (e.g., *poikki#telo* → *poikkitelo*).

D Human Agreement Study

We validate all three automated judges (LTE, Detection, Interpretation) against independent human annotation on the same three language pairs, Fi→En, Fr→En, and Ja→En. The human evaluation was conducted by senior Ph.D. students and research staff with Master’s degrees, all with expertise in the relevant languages.

D.1 LTE Judge and Prompt Selection

We first select the judge prompt by comparing candidate formulations against the human annotations with GPT-4o-mini, then, with that prompt fixed, compare GPT-4o-mini and GPT-5.2 against the three-annotator human majority on two directions (Fi→En, Fr→En). GPT-5.2 shows substantially higher agreement (Cohen’s κ =0.63 on Fi→En, 0.76 on Fr→En) than GPT-4o-mini (κ =0.46 on Fi→En, 0.37 on Fr→En) and is adopted as the LTE judge for all experiments.

D.2 LTE Agreement Results

Annotation setup. Each item shows the idiom, the source sentence, an automatically generated lit-

LTE Judge Prompt

System.
You are an expert in cross-lingual idiom translation.

Global Constraints
Output Format: Valid JSON object ONLY.
- Do NOT wrap in ```json or ```
- Do NOT add explanations before/after JSON

User.
Task
Evaluate two properties of the translation regarding the idiom.

Important
- The two judgments must be made independently.
- Both properties can be true or false independently.

Subjudgments (answer each independently and exact evidence from the translation)

Important distinction:
Literal mapping: the literal translation of words in the idiom (e.g., in the idiom ``drop the ball'', ``ball'' is translated literally as ``球'')
Intended meaning: the actual figurative meaning of the idiom (e.g., in the idiom ``drop the ball'', it is translated figuratively as ``犯错'')

Step 1: Extract the literal translation of each content word in the idiom.
Step 2: Literal Check - Does the translation preserve this literal mapping? (Yes / No)
Step 3: Meaning Check - Does the translation correctly express the intended meaning? (Yes / No)

Input
- Idiom: ``{idiom}''
- Intended meaning of the idiom: ``{meaning}''
- Source Sentence: ``{sentence}''
- Translation: ``{model_translation}''

Output Format (JSON only)

```
{
  ``idiom``: ``{idiom}``,
  ``intended_meaning``: ``{meaning}``,
  ``source_sentence``: ``{sentence}``,
  ``translation``: ``{model_translation}``,
  ``reasoning``: {
    ``explicit_literal_mapping``: ``Extract the literal translation of each content word in the idiom``,
    ``literal_check``: ``Yes/No + 10-20 word evidence``,
    ``meaning_check``: ``Yes/No + 10-20 word evidence``,
    ``confidence``: 0.0-1.0
  }
}
```

Detection Prompt

Read the following sentence carefully.

Sentence: {sentence}
Does this sentence contain an idiomatic expression (an expression whose meaning is different from the literal meaning of its individual words)?

If yes, identify the idiomatic expression and respond in JSON format:
{``has\_idiom``: true, ``idiom``: ``the idiomatic expression``}

If no idiomatic expression is found, respond:
{``has\_idiom``: false, ``idiom``: ````}

eral translation of the source sentence, the gold translation, the reference meaning, and the model

Detection Judge Prompt

System.
You are a linguistic judge. Determine whether two phrases refer to the same idiom, allowing for inflectional variations, pronoun changes, and partial forms. Output ONLY.
{`match`: true/false, `reason`: `brief reason`}

User.
Gold idiom: {gold_idiom}
Model detected: {detected_idiom}
Context: {sentence}

Do these refer to the same idiomatic expression? Consider:
(1) inflectional forms (drag one's feet = dragging my feet)
(2) pronoun substitution (one's = my/his/her/their)
(3) partial overlap (break the ice = broke the ice).
Reject completely different idioms. Output JSON only.

Interpretation Prompts

(a) What is the idiomatic meaning of `{idiom}` in: `{sentence}`? Output JSON only:
{`meaning`: `...`}

(b) Explain the idiomatic meaning of `{idiom}` in: `{sentence}`. Output JSON only:
{`meaning`: `...`}

(c) What does `{idiom}` idiomatically mean in: `{sentence}`? Output JSON only: {`meaning`: `...`}

Interpretation Judge Prompt

System.
You are a judge evaluating whether a model's explanation of an idiom is correct. Output ONLY.
{`correct`: true/false, `reason`: `brief reason`}

User.
Reference meaning: {reference_meaning}
Model's answer: {model_answer}

Does the model's answer correctly capture the essential meaning of the idiom? Be lenient: the answer need not be word-perfect, only semantically correct. Mark as incorrect if the model says `'I don't know'` or gives a wrong meaning. Output JSON only.

Reference Meaning Extraction Prompt

Provide the idiomatic meaning of the idiom `{idiom}` as used in the source sentence, based on the context of the sentence and its translation.

Source Sentence: {source}
Translation: {gold_translation}

Respond only with the idiom's meaning in English without any explanation, formatted as a JSON object:

```
{
  "output": "meaning content"
}
```

translation. Annotators mark the two conditions of the LTE definition in §3.1 separately, whether the idiom is rendered literally and whether its mean-

ing is conveyed, and the LTE label is computed from these two marks rather than assigned by hand. We collect LTE annotations from three indepen-

Dir.	Judge	Hum.	LLM	Agr.	κ
<i>(a) Majority-vote human vs. LLM judge</i>					
Fi→En	GPT-5.2	29.4	31.5	84.4	0.63
Fr→En	GPT-5.2	12.3	17.0	93.8	0.76
Ja→En	GPT-5.2	8.7	13.1	94.2	0.70
<i>(b) Inter-annotator agreement (3 annotators)</i>					
Fi→En	3-WAY FLEISS	$n=340$	72.9	0.58	
Fr→En	3-WAY FLEISS	$n=552$	93.1	0.78	
Ja→En	3-WAY FLEISS	$n=552$	84.9	0.47	

Table 7: Human evaluation of LTE across all four prompt conditions (all in %, except κ). **(a)** LTE rate and agreement between human judgment and the GPT-5.2 judge (raw agreement, Cohen’s κ), using the three-annotator majority vote. **(b)** Inter-annotator agreement among the three annotators (raw agreement, Fleiss’ κ).

dent annotators on all four prompt conditions of Fi→En ($n=340$), Fr→En ($n=552$), and Ja→En ($n=552$). Table 7(b) reports inter-annotator agreement on the final LTE label: it is substantial on Fr→En (Fleiss’ $\kappa=0.78$) and moderate on Fi→En ($\kappa=0.58$) and Ja→En ($\kappa=0.47$). Ja→En reaches 84.9% raw agreement, but LTE occurs in only 8.7% of its items. Because κ is deflated when one label is this rare, its lower value reflects the label distribution rather than annotator disagreement. Table 7(a) compares the three-annotator majority vote against GPT-5.2, our strongest judge: agreement is substantial across all three directions (Cohen’s $\kappa=0.76$ on Fr→En, 0.63 on Fi→En, and 0.70 on Ja→En), and the judge’s LTE rate closely tracks the human majority (17.0% vs. 12.3% on Fr→En; 31.5% vs. 29.4% on Fi→En; 13.1% vs. 8.7% on Ja→En). This confirms that the dual-condition LTE criterion is well-aligned with expert human judgment.

D.3 Detection and Interpretation Agreement Results

We collect Detection and Interpretation annotations from the same three annotators on all instances of Fi→En ($n=85$), Fr→En ($n=138$), and Ja→En ($n=138$). For Detection (Table 8), inter-annotator agreement ranges from 0.60 to 0.88 and majority-vote agreement with the GPT-4o-mini judge from 0.60 to 0.91. For Interpretation (Table 9), the corresponding ranges are 0.44 to 0.79 and 0.47 to 0.83.

E Linear Probe Implementation Details

All probes use scikit-learn’s LogisticRegression (solver lbfgs, L2 penalty, $C=1.0$, $\text{max_iter}=2000$, $\text{random_state}=42$). All other parameters are left

Dir.	Judge	Hum.	LLM	Agr.	κ
<i>(a) Majority-vote human vs. LLM judge</i>					
Fi→En	GPT-4o-mini	49.4	67.1	80.0	0.60
Fr→En	GPT-4o-mini	87.0	92.8	94.2	0.69
Ja→En	GPT-4o-mini	72.5	76.1	96.4	0.91
<i>(b) Inter-annotator agreement (3 annotators)</i>					
Fi→En	3-WAY FLEISS	$n=85$	82.4	0.77	
Fr→En	3-WAY FLEISS	$n=138$	84.8	0.60	
Ja→En	3-WAY FLEISS	$n=138$	92.8	0.88	

Table 8: Human evaluation of idiom Detection (all in %, except κ). **(a)** Detection accuracy and agreement between human judgment and the GPT-4o-mini judge (raw agreement, Cohen’s κ), using the three-annotator majority vote. **(b)** Inter-annotator agreement among the three annotators (raw agreement, Fleiss’ κ).

Dir.	Judge	Hum.	LLM	Agr.	κ
<i>(a) Majority-vote human vs. LLM judge</i>					
Fi→En	GPT-4o-mini	45.9	37.6	91.8	0.83
Fr→En	GPT-4o-mini	91.3	86.2	94.9	0.75
Ja→En	GPT-4o-mini	89.9	80.4	86.2	0.47
<i>(b) Inter-annotator agreement (3 annotators)</i>					
Fi→En	3-WAY FLEISS	$n=85$	84.5	0.79	
Fr→En	3-WAY FLEISS	$n=138$	91.3	0.64	
Ja→En	3-WAY FLEISS	$n=138$	81.8	0.44	

Table 9: Human evaluation of idiom Interpretation (any-hit across three probes, all in %, except κ). **(a)** Interpretation accuracy and agreement between human judgment and the GPT-4o-mini judge (raw agreement, Cohen’s κ), using the three-annotator majority vote. **(b)** Inter-annotator agreement among the three annotators (raw agreement, Fleiss’ κ).

at their scikit-learn defaults, and we perform no hyperparameter search. Features are the last input-token residual-stream hidden state at each of the 32 layers ($h_{i,\ell} \in \mathbb{R}^{4096}$ for Qwen3.5-9B), standardized with StandardScaler fit on the training partition only.

Naive random-fold probing (Table 3) uses 5-fold StratifiedKFold with shuffling. Because the two groups are imbalanced, we repeat it over five class-balanced subsampling seeds ($\{42, 123, 456, 789, 2026\}$), each time downsampling the majority class to the minority-class count. We report the mean $\pm$ SD of balanced accuracy across seeds, with ROC-AUC at the same layer.

Leakage-aware probing (Table 4) instead groups by idiom ID, using 5-fold StratifiedGroupKFold so that all prompt instances of an idiom stay in the same fold. Here we weight the classes rather than subsample, since down-sampling at the idiom level would leave too few groups to form five folds on

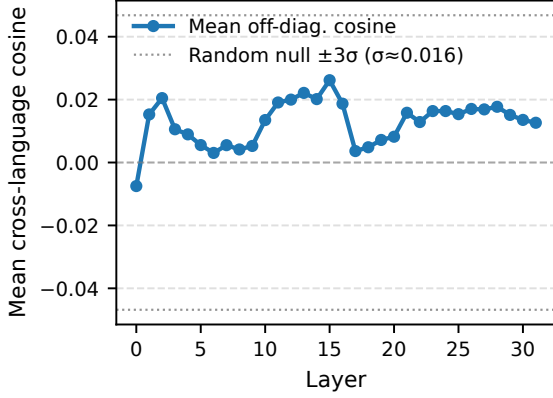


Figure 7: Mean cross-language LTD cosine by layer, averaged over the 15 pairwise comparisons among the six language pairs.

the smaller directions.

For the within-idiom permutation control (Table 4), we first subtract each idiom’s mean hidden state and probe the residuals under the same grouped scheme. The resulting accuracy is compared against the null distribution of accuracies obtained by shuffling the labels within each idiom, over 1,000 permutations. No direction exceeds its null (all $p \geq 0.17$).

F LTD Geometric Robustness Checks

Near-orthogonality is easy to over-read, so we check it four ways. In 4096 dimensions any two unrelated vectors are already close to orthogonal, so we first calibrate the observed cosines against random pairs. Even at chance level, the comparison so far uses only the peak probing layer, so we next repeat it at every depth. Even across all layers, directions can be pairwise orthogonal while still lying in a common low-dimensional subspace, a form of sharing that pairwise cosines cannot detect, so we then compare the subspaces themselves using principal angles. All of this assumes the directions are estimated stably enough to be compared, so we finally measure how much each language’s direction changes between adjacent layers.

Random-vector calibration. We sample 10^4 random unit-vector pairs in $\mathbb{R}^{4096}$ and take the range covering 99% of them, $|\cos| < 0.040$ (SD ≈ 0.016), as the range produced by unrelated vectors. The peak-layer off-diagonal LTD cosines in Figure 6 lie in $[-0.032, +0.044]$ with mean $|\cos| = 0.02$, inside this band. Only the largest pair slightly exceeds 0.040.

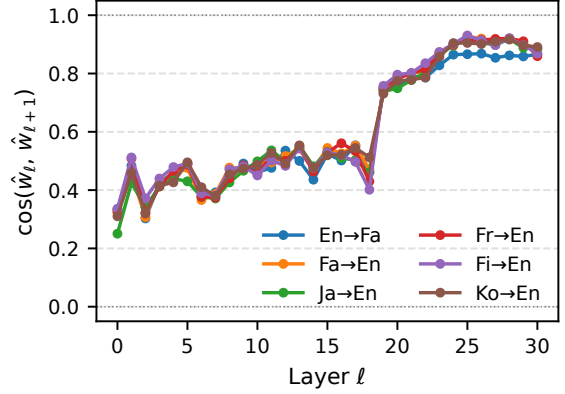


Figure 8: Within-language adjacent-layer stability $\cos(\hat{\mathbf{w}}_\ell, \hat{\mathbf{w}}_{\ell+1})$ for each language pair.

Per-layer cross-language cosine. At every layer, the mean cosine between LTDs from different language pairs stays within this same range, providing no evidence for a shared direction at any specific depth (see Figure 7).

Principal angles. We compute principal angles (Björck and Golub, 1973) between per-language LTD subspaces defined in two ways: (i) the top-5 layers by balanced probing accuracy (5-D per language), and (ii) all 32 layers (32-D per language). We compare each result against random subspaces of the same size. For the top-5 subspaces, none of the 15 language-pair comparisons yields a top-1 principal cosine more than 3σ above the random-subspace mean of 0.057. For the 32-layer subspaces, the mean top-1 principal cosine is 0.189, only slightly above the random-subspace value of 0.168. The remaining principal cosines fall off to the random-subspace level right away. If the languages shared a subspace of several dimensions, the first few would instead stay high, so we find no evidence for a shared cross-language subspace.

Within-language stability. Within each language pair, the direction becomes stable in the upper half of the model: $\cos(\hat{\mathbf{w}}_\ell, \hat{\mathbf{w}}_{\ell+1})$ rises from about 0.35 in early layers to 0.85 or above from L23 onward, and all six pairs follow the same curve (see Figure 8). The directions are therefore reproducible, so the near-orthogonality across languages is unlikely to be noise.

G Ablation Audit Details

To distinguish substantive output changes from evaluator-sensitive relabeling, the Fix/Harm counts in Table 5 include only paired samples where the

Pair	Idiom	Meaning	Change after LTD ablation	Type
En→Fa	<i>bite the dust</i>	fail or be defeated	زیر پا ماند ('remained under foot') → شکست خورد ('failed')	Real fix
En→Fa	<i>drop the ball</i>	make a mistake	توپ را زمین گذاشت ('put the ball down') → اشتباه بزرگی کرد ('made a big mistake')	Real fix
Ko→En	그림의 떡	unrealistic expectations	"cake in the picture" → "pie in the sky"	Real fix
Fi→En	<i>vetää turpa</i>	shut up	"To pull the tail?" → "Want to get your ass kicked?"	Changed, not fixed

Table 10: Representative audit cases after LTD ablation, shown with English glosses for readability. The table includes both real fixes and changed-but-not-fixed outputs.

translation text differs between the baseline and ablated generations. Table 10 shows representative cases drawn from this set, including both meaning-preserving fixes and changed-but-not-fixed outputs. The latter help calibrate whether the reduction reflects targeted intervention or broader output perturbation.

Consequences of Symmetric Re-scoring. Because LLM judgments are imperfect, we re-score both the baseline and the ablated output under the same protocol (§6.2) rather than treating Group A’s original labels as ground truth. This keeps an original judging error from counting as a Fix, and it lets ablation-induced harm be observed at all. The design has two consequences.

First, some instances receive a hedged verdict on `literal_check` or `meaning_check` (“Partial/No,” “Mostly Yes”) rather than a clean Yes/No. These cannot be assigned a binary LTE label, so we exclude them rather than force them into a class. This happens in only two directions, which is why n in Table 5 falls below Group A’s construction-time size in Table 3 there (36 vs. 42 and 204 vs. 220).

Second, re-scoring also means that some baseline outputs are no longer judged as LTE. These instances have no LTE left to fix, so ablation can only harm them, which is why Table 5 records Harm cases even though Group A is defined by initial LTEs.

H LTD Probing and Ablation Robustness Checks

The mechanistic analysis in §6.2 pools instances across all four prompt conditions, and the ablation in Table 5 removes the recovered direction at all 32 layers. We check both design choices. The Basic-only run repeats the probing and the ablation with

out cross-prompt pooling (§H.1), and the peak-layer run replaces the all-layer intervention with a more localized one at each direction’s peak probing layer (§H.2). Neither check changes the paper’s conclusions: probing behaves as the leakage account of Finding 6 predicts, and LTD ablation remains indistinguishable from a matched random-direction control. Confidence intervals quoted in the text are paired bootstrap intervals over 2 000 iterations.

H.1 Basic-only Prompt Robustness Check

Group A/B construction, layer-wise probing, LTD ablation, and the matched random-direction control are re-run restricted to the BASIC prompt condition alone. This shrinks Group A to $|A| = 63$ for Ja→En down to 4 for Fi→En (excluded as underpowered). All other pipeline details match §6.2.

On the only well-powered direction (Ja→En, $n=63$), LTD ablation and the random control move the LTE rate by almost the same amount (−11.1 pp vs. −9.5 pp), mirroring the pooled result in Table 5. In the smaller directions the LTD confidence interval likewise covers the random-direction mean. Restricting to BASIC therefore does not surface a prompt-specific effect that cross-prompt pooling had hidden. This restriction also independently corroborates the leakage account of Finding 6: with at most one instance per idiom, the dataset is idiom-partitioned by construction and naive K -fold accuracy is no longer inflated, consistent with Table 4.

Figures 9 and 10 report the Basic-only counterparts of Figures 5 and 7 (Fi→En excluded as underpowered). Both patterns match the pooled analysis.

H.2 Peak-Layer Robustness Check

The all-layer ablation in §6.2 is a broad intervention. We complement it with a single-layer analysis at each direction’s peak leakage-aware grouped-CV

	n	Δ LTE	Δ LTE (rand.)	Fix:Harm
En→Fa	16	−31.2 pp	−16.7 pp	5:0
Fr→En	17	−17.6 pp	−2.0 pp	4:1
Ko→En [†]	8	−25.0 pp	−20.8 pp	2:0
Fa→En [†]	13	−7.7 pp	−15.4 pp	2:1
Fi→En	—	$ A = 4$; excluded		
Ja→En	63	−11.1 pp [−20.6, −3.2]	−9.5 pp (± 4.8)	8:1

Table 11: Basic-only paired LTD ablation on Group A (Qwen3.5-9B). Δ LTE: change in LLM-judged LTE rate after ablation. Δ LTE (rand.): same for a norm-matched random direction (mean of three seeds). Fix:Harm: counts of initially-LTE cases corrected by ablation and initially non-LTE cases that become LTE. [†]Under-powered ($|A| < 20$).

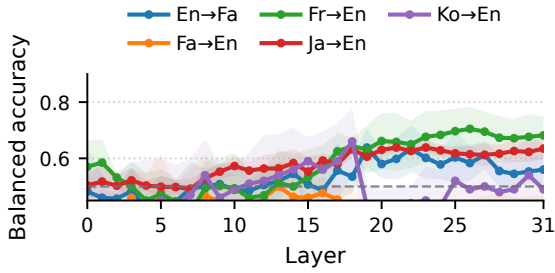


Figure 9: Basic-only all-layer linear probing curves for Group A vs. Group B on Qwen3.5-9B (compare Figure 5).

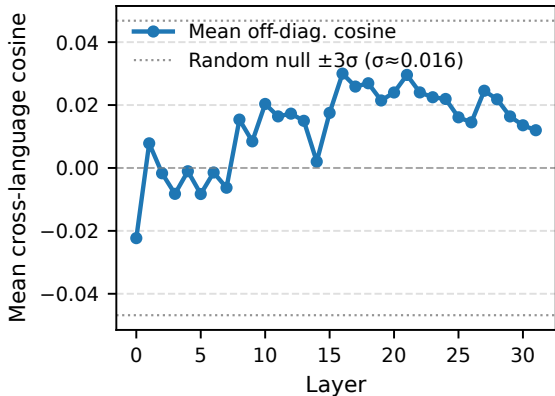


Figure 10: Basic-only mean cross-language LTD cosine by layer, averaged over the 10 pairwise comparisons among the five well-defined language pairs (compare Figure 7).

probing layer (Table 4), listed in the Peak L column below, with every other design choice matching Table 5 exactly. The LTD confidence interval covers the random-direction mean in *every* direction, including the only well-powered one (Ja→En, $|A|=220$: −10.0 pp vs. −6.8 pp), mirroring the pooled and Basic-only results (Appendix H.1). The peak-layer LTD effect is larger than the all-layer number in three directions and smaller in the other three, where in one case it even changes sign. Lo-

calization therefore does not recover a specificity gap that all-layer ablation was hiding. All-layer ablation remains the main analysis for the reasons given in §6.2, and the peak-layer results here are complementary rather than a replacement.

I Linguistic Analysis of LTB

We score each idiom with GPT-5.2 at temperature=0 and a 256-token output limit. One call rates one idiom, given the idiom and its reference meaning, and returns a JSON object with a 1–5 score and a short reason. Rating idiom compositionality is hard to standardize (Nordmann et al., 2014), and a bare 1–5 scale would leave each score point to the model’s own interpretation. The prompt therefore anchors the scale with three example idioms per source language, one each at scores 1, 3, and 5, all taken from outside the evaluation corpus. The prompt below shows the English anchors.

Idiom length is counted with each source language’s own Universal Dependencies tokenizer in stanza (Qi et al., 2020). Within each direction we take the median length over the instances entering the comparison, and call an instance short if its length is at or below that median and long otherwise. For Table 6 the medians are 2 for Fi→En and Ko→En, 3 for Ja→En, 4 for En→Fa and Fa→En, and 4.5 for Fr→En.

J Linguistic Analysis of Full Data

Finding 9 restricts the comparison to the LTB-conditional subset ($D=1, I=1$). Here we repeat it on the full evaluation set, without conditioning on detection and interpretation. The pattern is the same: low-compositionality idioms have higher LTE rates than high-compositionality ones in five of six directions (Table 13). Idiom length again shows no consistent effect.

	Peak L	<i>n</i>	Δ LTE	Δ LTE (rand.)	Fix:Harm
En→Fa	15	42	−14.3 pp	−10.3 pp	6:0
Fr→En	31	49	+8.2 pp	+2.7 pp	3:7
Ko→En [†]	11	17	−11.8 pp	−11.8 pp	2:0
Fa→En	16	41	−19.5 pp	−14.6 pp	9:1
Fi→En [†]	17	13	0.0 pp	−7.7 pp	3:3
Ja→En	11	220	−10.0 pp	−6.8 pp	32:10

Table 12: Peak-layer paired LTD ablation on Group A (Qwen3.5-9B, mixed prompts). Peak L: the leakage-aware grouped-CV peak layer per direction, at which both the LTD and the random direction are ablated. Δ LTE: change in LLM-judged LTE rate after ablation. Δ LTE (rand.): same for a norm-matched random direction (mean of three seeds). Fix:Harm: counts of initially-LTE cases corrected by ablation and initially non-LTE cases that become LTE. [†]Under-powered ($|A| < 20$).

Compositionality Scoring Prompt

You are a linguist evaluating the compositionality of idiomatic expressions.

Compositionality is the degree to which an idiom’s figurative meaning can be inferred from the literal meanings of its component words.

Rating scale:

- 1 (Fully opaque): The meaning cannot be inferred at all from the individual words.
- 2 (Mostly opaque): Very little of the meaning can be guessed from the components.
- 3 (Semi-compositional): Some aspects of the meaning can be inferred through metaphorical reasoning.
- 4 (Mostly compositional): The meaning is largely inferable from the components.
- 5 (Fully compositional): The meaning is directly derivable from the individual words.

Here are calibration examples:

Score 1: ‘‘fly off the handle’’ = ‘‘to suddenly become very angry’’
Reason: ‘‘fly’’, ‘‘off’’, ‘‘handle’’ have no connection to ‘‘anger’’

Score 3: ‘‘break the ice’’ = ‘‘to relieve tension in a social situation’’
Reason: ‘‘break’’ suggests disrupting, ‘‘ice’’ suggests coldness/stiffness --- requires metaphorical reasoning

Score 5: ‘‘break one’s heart’’ = ‘‘to cause deep emotional pain’’
Reason: ‘‘break’’ directly maps to damage, ‘‘heart’’ directly maps to emotions

Now rate this idiom:
Idiom: {idiom}
Meaning: {meaning}

Rate the compositionality of this idiom on a 1-5 scale.
Output ONLY a JSON object:
{‘‘compositionality_score’’: <1-5>, ‘‘reason’’: ‘‘brief reason in 10-20 words’’}

Direction	Low comp.	High comp.	Short	Long
En→Fa	7.8%	5.5%	6.3%	6.1%
Fr→En	18.7%	11.3%	15.7%	18.4%
Ko→En	6.5%	3.2%	4.1%	2.1%
Fa→En	18.5%	14.3%	15.7%	18.8%
Fi→En	31.4%	31.7%	32.1%	30.5%
Ja→En	17.4%	11.5%	13.0%	13.7%

Table 13: LTE rate by compositionality and idiom length on the full evaluation set. Low/high comp.: scores 1–3 vs. 4–5. Short/long: at or below vs. above each direction’s median length.

K Results for Remaining Models

Figures 11–18 provide additional behavioral results for the other three Qwen models and Llama-3.3-70B-Instruct. Findings 1–5 remain consistent across model sizes and families.

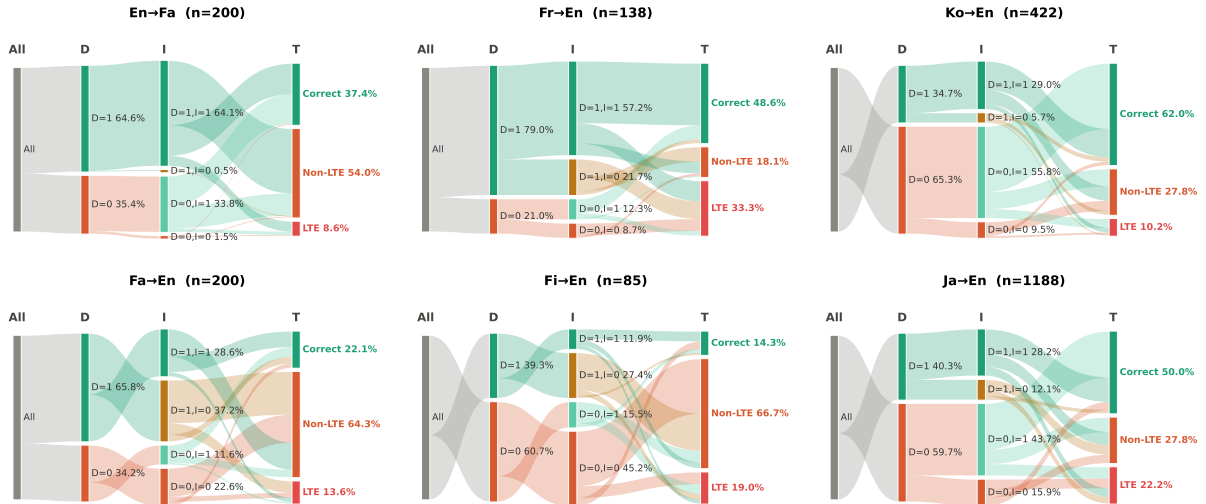


Figure 11: Sankey diagram of IdiomDIT outcomes for all six language pairs (Qwen3-4B, BASIC prompt).

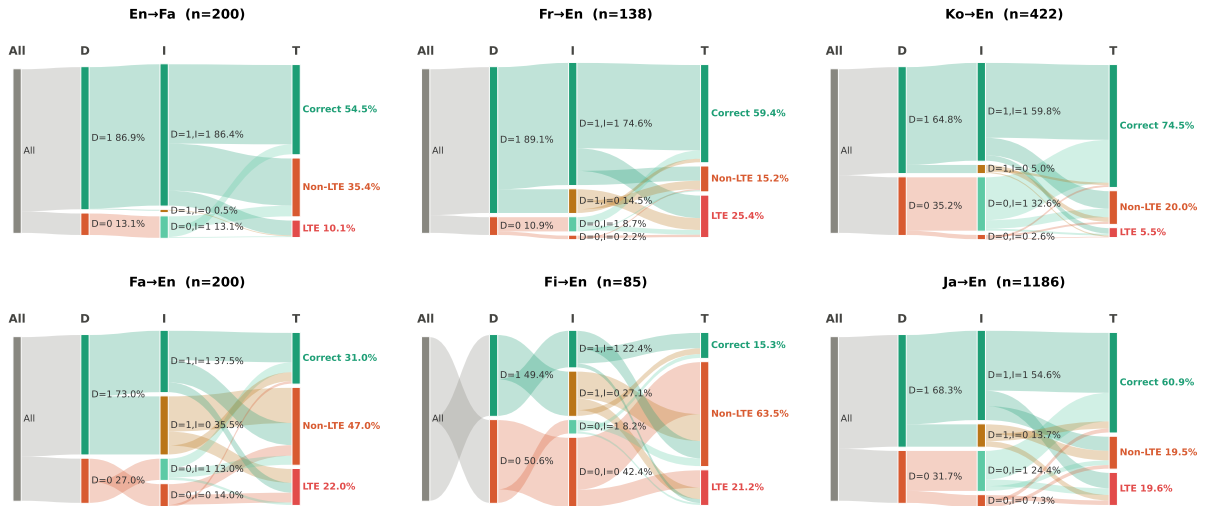


Figure 12: Sankey diagram of IdiomDIT outcomes for all six language pairs (Qwen3-8B, BASIC prompt).

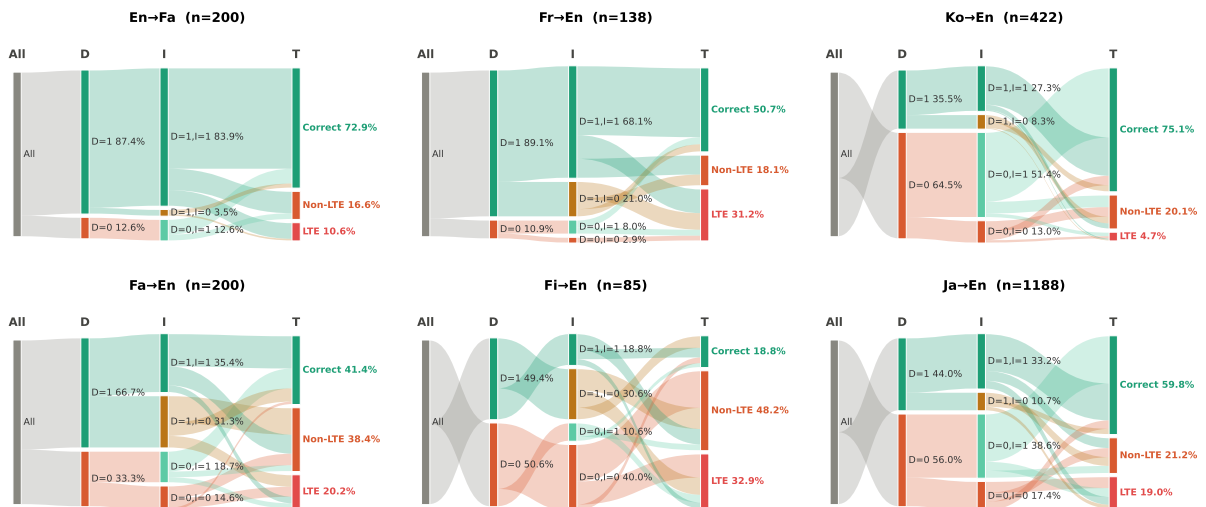


Figure 13: Sankey diagram of IdiomDIT outcomes for all six language pairs (Qwen3.5-4B, BASIC prompt).

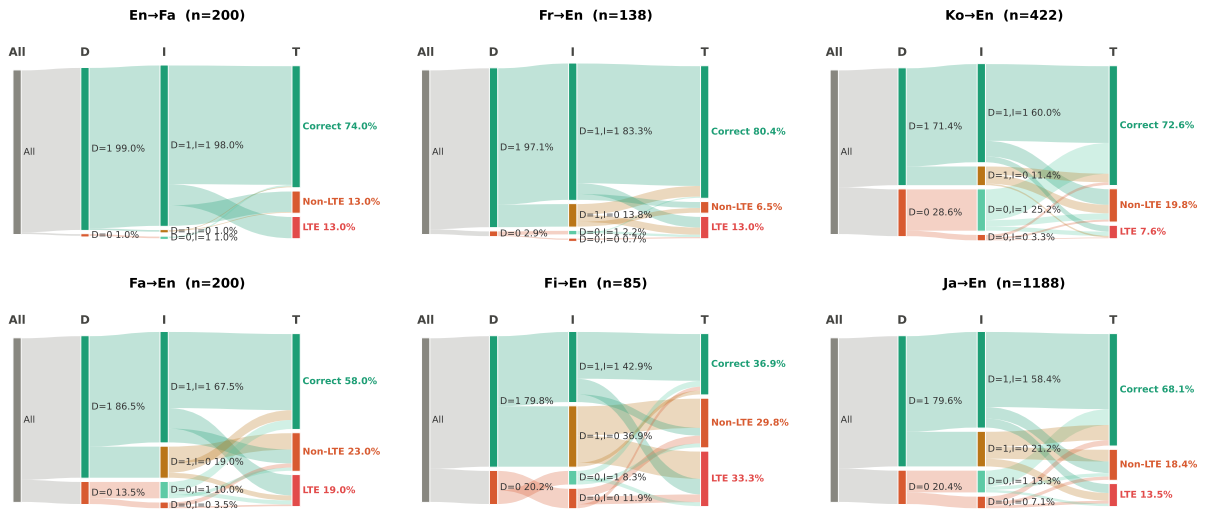


Figure 14: Sankey diagram of IdiomDIT outcomes for all six language pairs (Llama-3.3-70B-Instruct, Basic prompt).

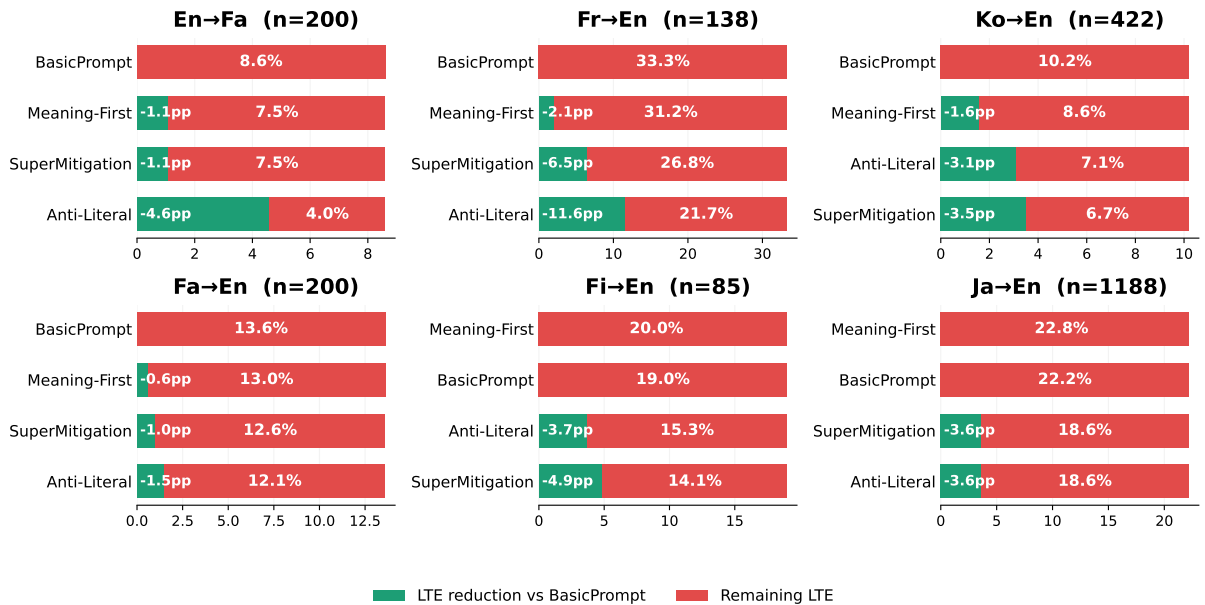


Figure 15: Prompt mitigation effect on LTE rate (Qwen3-4B, all six language pairs).

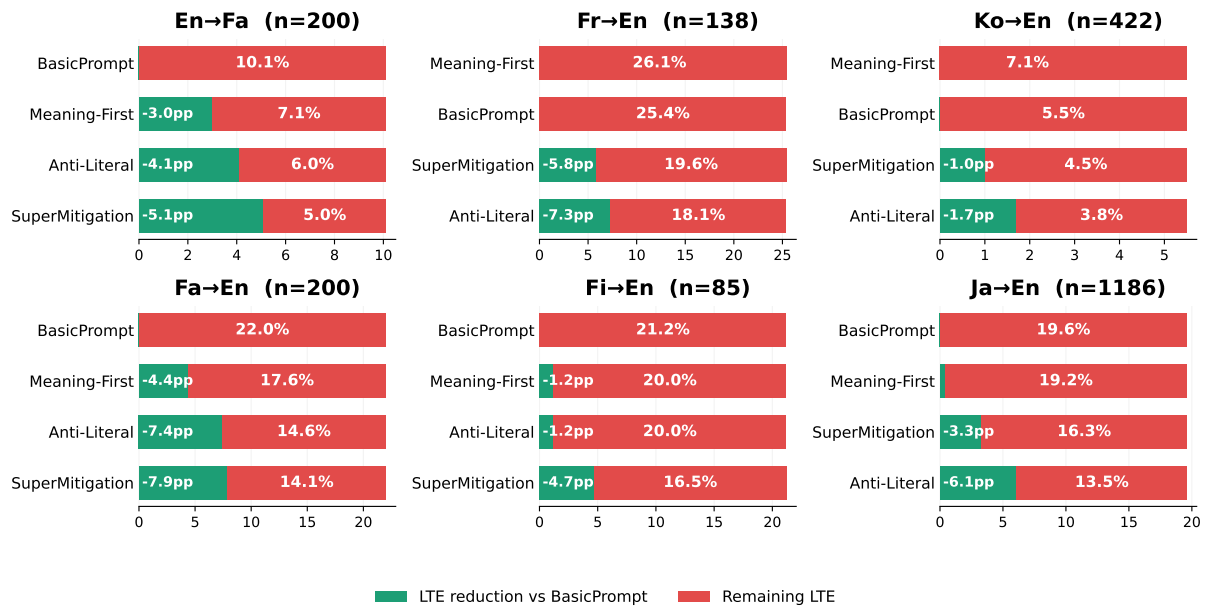


Figure 16: Prompt mitigation effect on LTE rate (Qwen3-8B, all six language pairs).

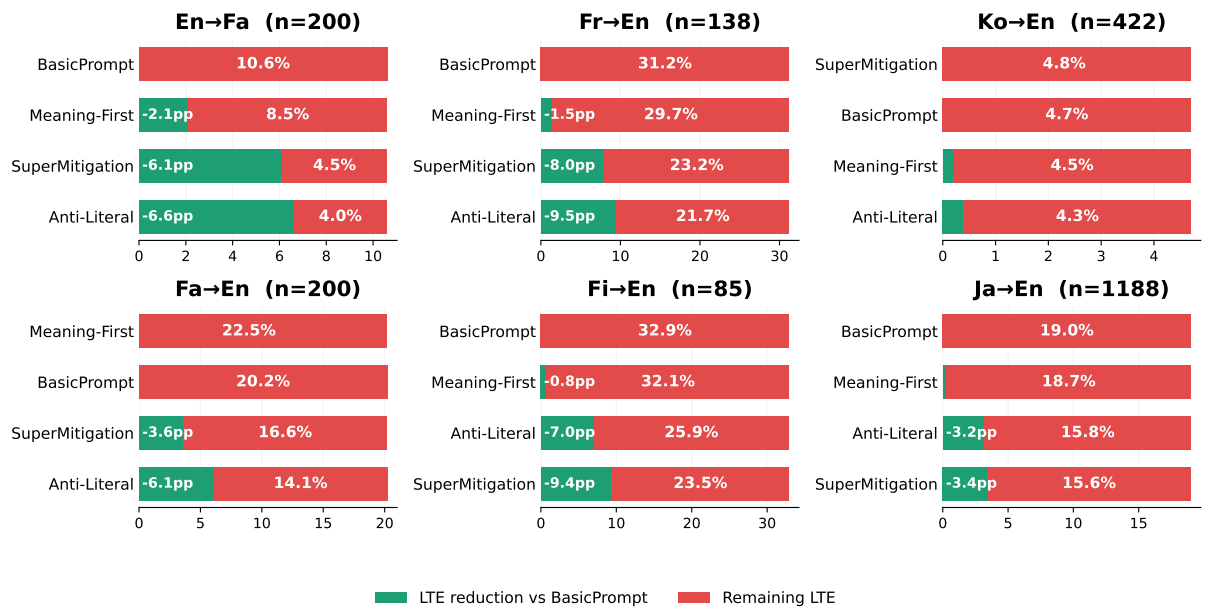


Figure 17: Prompt mitigation effect on LTE rate (Qwen3.5-4B, all six language pairs).

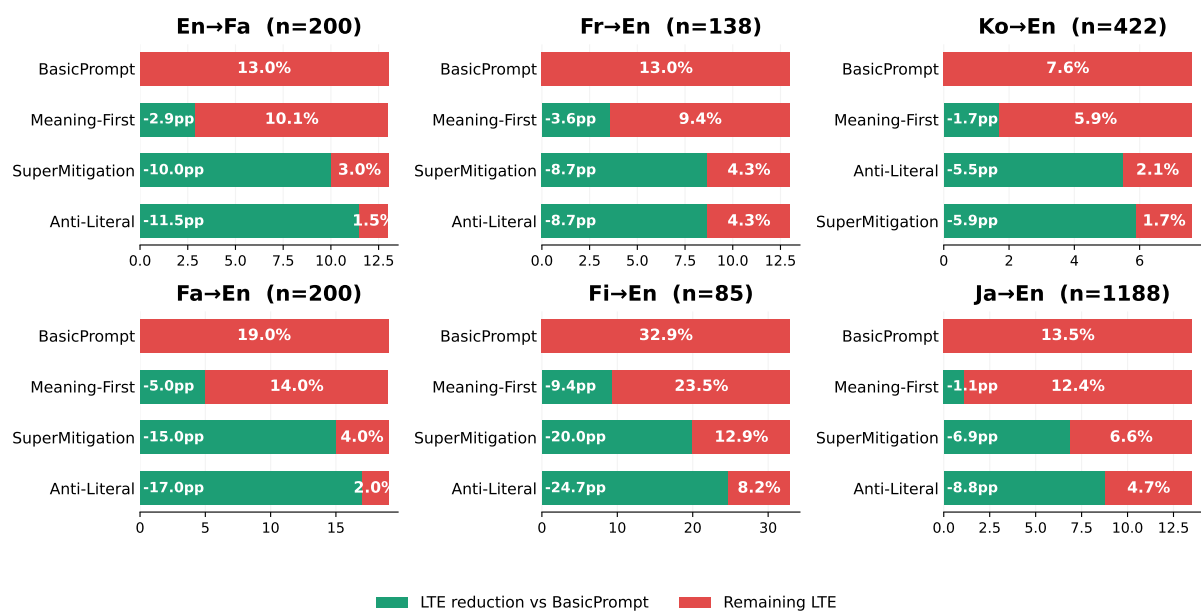


Figure 18: Prompt mitigation effect on LTE rate (Llama-3.3-70B-Instruct, all six language pairs).