HOMURA: Taming the Sand-Glass for Time-Constrained LLM Translation via Reinforcement Learning

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have achieved remarkable strides in multilingual translation but are hindered by a systemic cross-lingual verbosity bias, rendering them unsuitable for strict time-constrained tasks like subtitling and dubbing. Current prompt-engineering approaches struggle to resolve this conflict between semantic fidelity and rigid temporal feasibility. To bridge this gap, we first introduce Sand-Glass, a benchmark specifically designed to evaluate translation under syllable-level duration constraints. Furthermore, we propose HOMURA, a reinforcement learning framework that explicitly optimizes the trade-off between semantic preservation and temporal compliance. By employing a KL-regularized objective with a novel dynamic syllable-ratio reward, HOMURA effectively "tames" the output length. Experimental results demonstrate that our method significantly outperforms strong LLM baselines, achieving precise length control that respects linguistic density hierarchies without compromising semantic adequacy.

1 Introduction

With the rapid advancement of Large Language Models (LLMs), their multilingual capabilities have achieved unprecedented semantic adequacy and fluency in Neural Machine Translation (NMT) (Hendy et al., 2023; Jiao et al., 2023; Zhu et al., 2024). However, this linguistic prowess is accompanied by a systemic cross-lingual verbosity bias (Briakou et al., 2024; Manakhimova et al., 2024). As illustrated in Figure 1, LLM-generated translations tend to be significantly longer than their source utterances, a "chattiness" that persists even when semantic content remains constant. While acceptable in general text generation, this expansion becomes a critical failure mode in time-constrained scenarios such as subtitles, video dubbing, and simultaneous interpretation,where translation length constitutes a rigid temporal budget

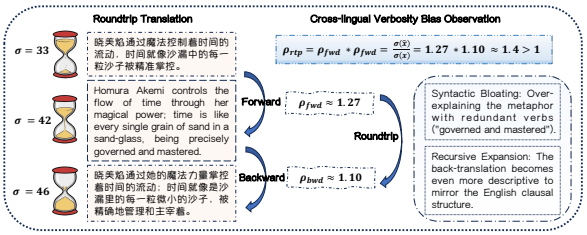


Figure 1: Illustration of cross-lingual verbosity bias. Unlike forward metrics, the **Roundtrip Expansion Ratio** ρ_{rtp} isolates model-induced redundancy from linguistic density shifts. A value of $\rho_{\text{rtp}} > 1.0$ indicates systemic inflation despite constant semantic content. Formal definition and diagnostic study are provided in Section 2.

that strictly limits the output duration (Karakanta et al., 2020b; Jørgensen and Mengshoel, 2025; Gaido et al., 2024). To operationalize the temporal budget, we approximate the available speaking time with a cross-lingually comparable syllable budget.

Viewed through an information-theoretic lens (Shannon, 1948; Davisson, 1972), these temporal budgets act as a strict rate constraint: with a limited number of target syllables available, the model must pack maximal meaning into minimal length, implicitly operating near a rate-distortion limit on the achievable fidelity compression trade-off. In our setting, the rate is the syllable budget, and distortion can be understood as the loss in meaning or adequacy relative. Unfortunately, off-the-shelf LLMs struggle to respect these limits. This failure is deeply rooted in the post-training alignment process, where learned reward models frequently exhibit a "length bias" that conflates verbosity with quality (Saito et al., 2023; Singhal et al., 2023). Furthermore, cross-lingual typological differences systematically skew length ratios, transforming length control from a superficial formatting task into a first-class modeling challenge. [1]

[1] Detailed related works can be found in appendix A.

To address this fundamental mismatch between LLM capabilities and temporal constraints, we argue that length control should be treated as an optimization objective rather than a post-hoc adjustment. To this end, we first operationalize the problem by constructing *Sand-Glass*, a dedicated benchmark for time-constrained media localization. The name reflects the rigid temporal budget that "drains" as the model produces redundant tokens. Unlike generic translation datasets, Sand-Glass incorporates syllable-based duration proxies calibrated by Information Density (ID), derived from real-world colloquial subtitles enabling precise evaluation of length compliance.

Building on this testbed, we propose *HOMURA* (**H**ard **O**ptimization for **M**ultilingual **U**tterance **R**eduction and **A**lignment), a reinforcement learning framework designed to "tame the sandglass" by explicitly regulating translation length without requiring supervised compression datasets. Instead of relying on brittle prompts, HOMURA optimizes the quality compression trade-off directly. We design a novel reference free reward function that combines a dynamic syllable ratio penalty with a semantic fidelity reward, encouraging the model to condense information density without sacrificing core meaning.

Our core contributions are as follows:

- **Diagnostic Analysis:** We systematically quantify the verbosity bias in LLM-based translation under temporal constraints, providing cross-lingual diagnostics (e.g., the ρ_{rtp} metric) that reveal the limitations of unconstrained models as visualized in Figure 1.
- **Benchmark Construction:** We introduce Sand-Glass, a specialized dataset tailored for time-constrained translation, featuring syllable-level duration budgets to simulate real-world dubbing and subtitling scenarios.
- **Methodological Innovation:** We propose HOMURA, a KL-regularized RL framework that dynamically balances strict length constraints with semantic preservation, significantly outperforming strong LLM baselines in both compliance and quality.

2 Pilot Study: Benchmark and Verbosity Bias

2.1 The Sand-Glass Benchmark

To diagnose cross-lingual verbosity bias under temporal constraints, we introduce Sand-Glass, a benchmark derived from colloquial video transcripts. Unlike generic datasets, it incorporates syllable-based duration proxies to strictly test time-sensitive compliance. The construction involves duration-aware segmentation based on speech pauses and core event extraction for semantic evaluation. We refer readers to Appendix D for detailed protocols on dataset filtering and event extraction.

Language	**Syllable Rate** (σ/s)	**Info. Density** (Normalized)	**Expansion** (from Zh)
Mandarin (Zh)	5.18 ± 0.15	0.94	$1.00\times$
English (En)	6.19 ± 0.16	0.91	$\approx 1.03\times$
German (De)	5.97 ± 0.19	0.79	$\approx 1.19\times$
Spanish (Es)	7.82 ± 0.16	0.63	$\approx 1.49\times$

Table 1: Cross-linguistic statistics adapted from Pellegrino et al. (2011). Lower density necessitates higher expansion factors when translating from Mandarin.

2.2 Syllable as Duration Proxy

Our use of syllables as duration proxies is grounded in the Iso-Information Principle. Pellegrino et al. (2011) demonstrated that while Syllable Rate (SR) vary drastically across languages (e.g., Spanish $7.82\sigma/s$ [syllables/s] vs. Mandarin $5.18\sigma/s$), the universal Information Rate remains nearly constant at ≈ 39 bits/s (Coupé et al., 2019).Recent findings further corroborate this by quantifying a universal surprisal–duration trade-off (Pimentel et al., 2021), showing that languages systematically adjust segment duration to smooth information transmission.

As shown in Table 1, this trade-off implies that low-density languages must utilize higher syllable counts to convey equivalent content. Translating from a high-density source (Mandarin, ID=0.94) to a lower-density target (e.g., Spanish, ID=0.63) necessitates a *linguistically required expansion*. A direct 1:1 mapping would result in $\approx 33\%$ information loss (derived from the density ratio $\frac{0.63}{0.94}$). Thus, our constraints are non-uniform and calibrated to these density ratios to distinguish legitimate expansion from model verbosity.

2.3 Quantifying Cross-Lingual Verbosity Bias

Standard metrics like the Forward Expansion Ratio defined as $\rho_{fwd} = \sigma(y)/\sigma(x)$ for a source x and its translation y often confound *linguistic necessity* with *model hallucinations*. For instance, a high ρ_{fwd} in Spanish might simply reflect its naturally lower information density rather than model-induced redundancy.

To isolate systemic bias, we introduce the *Roundtrip Expansion Ratio* (ρ_{rtp}). By translating

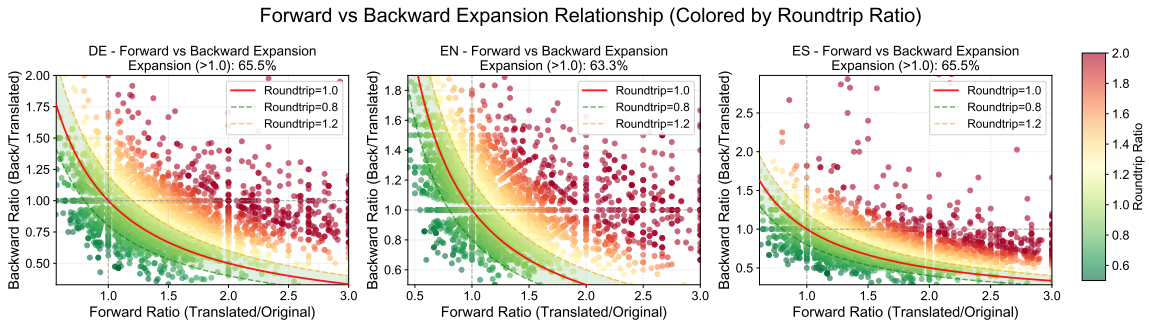


Figure 2: Forward vs. backward expansion relationship (colored by round-trip ratio ρ_{rtp}). The red curve represents the unity baseline $\rho_{\text{rtp}} = 1.0$ ($y = 1/x$). Points above this curve indicate systemic model-induced length inflation. Unconstrained models show a median $\rho_{\text{rtp}} > 1.10$.

the target output y back into the source language ($\hat{x}$) and comparing its syllable count to the original input x, we define:

$$\rho_{rtp} = \rho_{fwd} \cdot \rho_{bwd} = \frac{\sigma(y)}{\sigma(x)} \cdot \frac{\sigma(\hat{x})}{\sigma(y)} = \frac{\sigma(\hat{x})}{\sigma(x)}, \quad (1)$$

where $\sigma(\cdot)$ denotes the syllable count of a sequence. Theoretically, since this comparison is performed within the same source language, ρ_{rtp} should gravitate towards 1.0. Any value where $\rho_{rtp} > 1.0$ serves as a density-invariant indicator of model-induced inflation.

Empirical analysis on Sand-Glass (Figure 2) confirms that this verbosity is systemic rather than anecdotal. By plotting the joint distribution of expansion ratios, we observe a pervasive upward shift from the unity baseline ($\rho_{rtp} = 1.0$), with the vast majority of segments clustered in the "inflation zone." This concentration demonstrates a consistent model-induced redundancy that persists across various frontier models and language pairs (see Appendix L for statistical breakdowns).

Calibrated Target Bounds. To bridge this gap, we define Target Syllable Bounds ($\mathcal{B}_L$) by contrasting model inflation (μ_{LLM}) against Theoretical Expansion (μ_{theo}). As shown in Table 2, LLMs produce significant surplus bias (Δ_{bias}). We establish $\mathcal{B}_L$ to enforce a strict condensation regime. Notably, we lower these targets beyond theoretical baselines to address real-world dubbing and subtitling complexities, specifically to buffer against rapid speech and synchronization constraints. Accordingly, we configure the bounds as follows: English ($\mathcal{B}_{En} \in [0.8, 0.9]$) is set below parity to force active summarization; German ($\mathcal{B}_{De} \in [0.9, 1.0]$) maintains relative pressure despite lower density; and Spanish ($\mathcal{B}_{Es} \in [1.0, 1.1]$) significantly reduces the unconstrained baseline of $1.84\times$ to ensure intelligibility under strict time limits while respecting its information density.

Target Lang	Theoretical (μ_{theo})	LLM Baseline (μ_{LLM})	Bias (Δ)	Ours (Target) ($\mathcal{B}_L$)
En	$1.03\times$	$1.35\times$	$+0.32$	$\mathbf{0.8 \sim 0.9}$
De	$1.19\times$	$1.59\times$	$+0.40$	$\mathbf{0.9 \sim 1.0}$
Es	$1.49\times$	$1.84\times$	$+0.35$	$\mathbf{1.0 \sim 1.1}$

Table 2: Comparison of expansion ratios. Target Bounds ($\mathcal{B}_L$) are set to remove model bias (Δ) and enforce semantic condensation.

3 Method

We formulate syllable-level control as a standard controllable text generation task(Li et al., 2024; Gu et al., 2025), where our objective is to maximize translation quality while satisfying the prescribed syllable-ratio constraint. Our method is illustrated in Figure 3.

3.1 KL-Regularized Objective for Controllable Generation

We study conditional generation with an input x and a control condition c. Our objective is to learn a policy $\pi_\theta(y \mid x, c)$ that satisfies c.

Target Distribution via a Controlled Posterior. A natural target distribution for controllable generation is the controlled posterior

$$P(y \mid x, c) \propto P(c \mid x, y)\, P_0(y \mid x), \quad (2)$$

where $P(c \mid x, y)$ measures how well an output y meets the control condition, and $P_0(y \mid x)$ denotes a fixed reference model that captures the unconstrained translation distribution given the input x.

KL Projection. We fit $\pi_\theta(\cdot \mid x, c)$ by projecting it onto the target distribution:

$$\pi_\theta^\star(\cdot \mid x, c) = \arg\min_{\pi_\theta} \mathrm{KL}(\pi_\theta(\cdot \mid x, c) \,\|\, P(\cdot \mid x, c)). \quad (3)$$

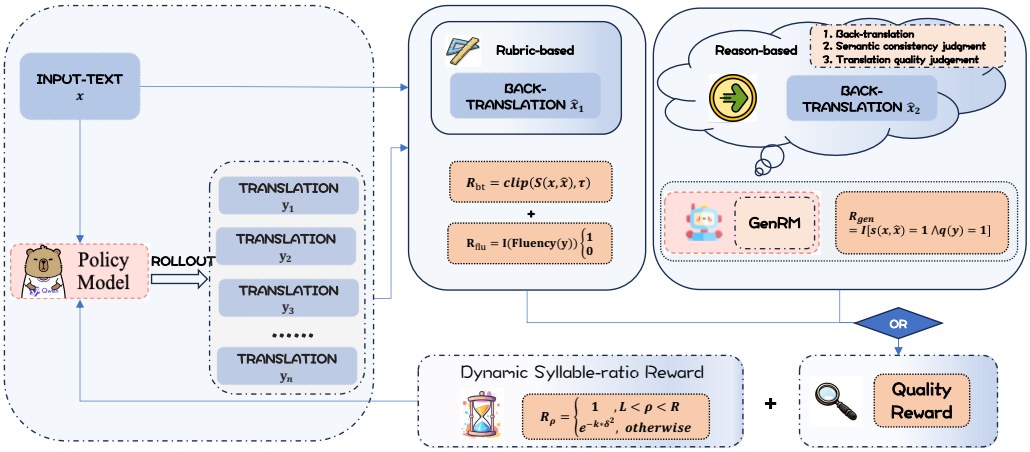


Figure 3: Overview of our HOMURA framework. Given a source utterance x and a syllable-ratio budget c, we optimize a KL-regularized policy $\pi_\theta(y \mid x, c)$ with GRPO. The reward combines a length-compliance term R_{len} and a quality term R_{qual} (rubric-based or reason-based), encouraging concise yet faithful translations.

Ignoring constants independent of θ and using Eq. (2), this objective is equivalent to maximizing

$$J(\theta) = \mathbb{E}_{y \sim \pi_\theta}\left[\log P(c \mid x, y) - \beta \log \frac{\pi_\theta(y|x,c)}{P_0(y|x)}\right], \tag{4}$$

where $\beta > 0$ is a temperature (or weighting) coefficient that controls the strength of KL regularization. In practice, we model $\log P(c \mid x, y)$ with a task reward $R(x, c, y)$.

Eq. (4) can be viewed as a KL-regularized reinforcement learning objective: $R(x, c, y)$ encourages satisfying the control condition, while the KL term discourages drifting away from the base model P_0. The coefficient β controls the trade-off between control strength and faithfulness.

Optimization with GRPO. We optimize Eq. (4) using GRPO (Shao et al., 2024), a lightweight PPO-style algorithm for sequence-level rewards. For each (x, c), we sample multiple candidates from the previous policy $\pi_{\theta_{\text{old}}}$, compute rewards and group-normalized advantages, and update π_θ with a clipped policy objective together with the KL regularization toward P_0.

3.2 Reward Design

In our setting, the control condition c corresponds to satisfying a strict length budget while maintaining semantic consistency with the source. This requires a reward that simultaneously encourages length compliance and discourages meaning drift. We interpret $\log P(c \mid x, y)$ in Eq. (4) as a task reward, denoted by $R(x, c, y)$, and instantiate it as:

$$R(x, c, y) = \lambda_{\text{len}}\, R_{\text{len}}(x, y) + \lambda_{\text{qual}}\, R_{\text{qual}}(x, y), \tag{5}$$

where x is the source utterance, y is the generated translation, and $\lambda_{\text{len}}, \lambda_{\text{qual}} \geq 0$ control the trade-off between temporal alignment and semantic fidelity.

3.2.1 Dynamic Syllable-ratio Reward

To ensure the translation length aligns naturally with the source speech across varying scales, we propose a dynamic syllable-ratio reward. The syllable ratio ρ for a source-target pair (x, y) is defined as:

$$\rho(x, y) = \frac{\sigma(y)}{\max(\sigma(x), 1)}, \tag{6}$$

where $\sigma(\cdot)$ is the syllable count of a sequence.

Recognizing that strict length constraints must be sensitive to both linguistic density and utterance length, we implement a dynamic acceptance interval $[L(x), R(x)]$. We initialize the baseline bounds $[L_0, R_0]$ using the language-specific Target Bounds $\mathcal{B}_L$ set in Section 2.3. This ensures the optimization target respects the cross-lingual density hierarchy established in our theoretical analysis.

However, for shorter utterances, the "granularity" of syllable counting introduces high variance (discretization noise), making rigid bounds brittle. To address this, for utterances where the source length $\sigma(x)$ is below the corpus mean μ, we define a scaling factor $\gamma(x)$ to adaptively relax the lower threshold:

$$\gamma(x) = \alpha_1 + \alpha_2 \left(\frac{\sigma(x)}{\mu}\right)^{1/2}, \tag{7}$$

where α_1, α_2 are hyperparameters governing the degree of relaxation. The final dynamic bounds are formulated as $[L(x), R(x)] = [L_0 \cdot \gamma(x),\ R_0]$.

To facilitate stable policy optimization, we avoid sparse or binary signals in favor of a smoother reward landscape. The reward function $R_{\text{len}}(x, y)$ evaluates compliance with these dynamic boundaries via a squared exponential decay:

$$R_\rho = \begin{cases} 1 & \text{if } \rho \in [L(x), R(x)] \\ \exp(-k\delta^2) & \text{otherwise} \end{cases} \quad (8)$$

where $\delta(x, y) = \max\big(|\rho(x, y) - L(x)|, |\rho(x, y) - R(x)|\big)$ denotes the deviation. This non-linear formulation provides a soft penalty that effectively guides the model toward the density-calibrated region while preventing training instability associated with hard-truncation rewards. In Appendix F, we further investigate the effectiveness of a static syllable-ratio reward.

3.2.2 Rubric-based Quality Reward

We adopt a two-part quality rubric(Hashemi et al., 2024; Liu et al., 2023), covering (i) semantic fidelity and (ii) linguistic well-formedness, and implement it as the following reward components.

Semantic Fidelity Reward Inspired by the probabilistic duality in (He et al., 2016), we enforce content preservation via a reconstruction-based reward. We employ a frozen `DeepSeek-V3` (DeepSeek-AI, 2024) to back-translate the hypothesis y into $\hat{x} = f_{bt}(y)$. To measure semantic retention, we compute the cosine similarity $S(x, \hat{x})$ between embeddings of the source x and reconstruction $\hat{x}$ extracted by `Qwen-Embedding-0.6B` (Zhang et al., 2025):

$$S(x, \hat{x}) = \frac{\mathbf{v}_x \cdot \mathbf{v}_{\hat{x}}}{\|\mathbf{v}_x\| \|\mathbf{v}_{\hat{x}}\|} \quad (9)$$

The final reward is clipped to stabilize optimization:

$$R_{bt}(x, y) = \text{clip}(S(x, \hat{x}), \tau_{min}, \tau_{max}) \quad (10)$$

Linguistic Well-formedness Reward To prevent the telegraphic outputs often associated with aggressive compression, we explicitly penalize linguistic degradation. We utilize a frozen `DeepSeek-V3` as a judge to evaluate translations against a rubric covering: (i) *grammaticality*, (ii) *readability*, and (iii) *language consistency* (e.g., no code-switching). The model outputs a binary decision $s_{flu} \in \{0, 1\}$, serving directly as the reward:

$$R_{flu}(x, y) = s_{flu} \quad (11)$$

Full prompting details are in Appendix I.

3.2.3 Reason-based Quality Reward

Beyond the simple compositional rewards, we explore an end-to-end generative reward model (GenRM)(Mahan et al., 2024; Liang et al., 2025) that evaluates translation quality via explicit reasoning. GenRM is trained using structured Chain-of-Thought (CoT) annotations synthesized by `Gemini-2.5-Pro` (Comanici et al., 2025). To ensure high-quality reasoning, we filter the raw CoT data using the aforementioned rubric-based method. The model is instantiated from `Qwen3-32B` (Yang et al., 2025) and optimized via supervised fine-tuning (SFT) over the full CoT sequence and the final decision token.

Given a source sentence x and a translation hypothesis y, GenRM performs a single long CoT reasoning process that can be abstracted as

$$\hat{x} = f_{\text{bt}}(y),\ s(x, \hat{x}), q(y) \in \{0, 1\}, \quad (12)$$

where f_{bt} denotes implicit back-translation, $s(x, \hat{x})$ indicates whether the core information in x is preserved, and $q(y)$ evaluates the intrinsic quality of y in terms of adequacy and fluency. Based on these decisions, GenRM outputs a binary quality reward

$$R_{gen}(x, y) = \mathbb{I}\big[s(x, \hat{x}) = 1 \wedge q(y) = 1\big]. \quad (13)$$

This formulation yields a reference-free, interpretable quality signal and enforces an explicit conjunction between information preservation and translation quality. Concrete examples of the CoT reasoning are provided in the Appendix H.

4 Experiments and Results

4.1 Experimental Setup

We evaluate SOTA LLMs under five strategy categories that span the length-control spectrum:

- **Category 1: Unconstrained LLMs**. Frontier models (`GPT-5` (OpenAI, 2025), `Claude-4.1-Opus` (Anthropic, 2025), `Gemini-2.5-Pro`, `DeepSeek-V3`) with standard decoding. For `GPT-5` and `Gemini-2.5-Pro`, we set low reasoning effort to approximate direct translation. These serve as semantic upper bounds.

Table 3: Performance Comparison on Sand-Glass Across All Language Pairs (BLEU-ρ, Cometkiwi, BT-CERR). All results are averaged by 3 times.

Model	Evaluation Metrics on Sand-Glass														
	Zh → En					Zh → De					Zh → Es				
	IB	Cometkiwi	BT-CERR	BLEU-ρ	AVG-Tokens	IB	Cometkiwi	BT-CERR	BLEU-ρ	AVG-Tokens	IB	Cometkiwi	BT-CERR	BLEU-ρ	AVG-Tokens
Category 1: w/o Syllable Constraint															
`claude-4.1-opus`	×	0.741	0.943	0.313	17.1	×	0.676	0.882	0.247	24.2	×	0.668	0.900	0.220	21.3
`deepseek-v3`	×	0.748	0.941	0.287	46.3	×	0.676	0.889	0.222	62.3	×	0.671	0.884	0.191	99.6
`gemini-2.5-pro`*	×	0.739	0.941	0.307	12.1 (916.9)	×	0.666	0.877	0.238	14.2 (125.8)	×	0.668	0.883	0.212	14.0 (1532.7)
`gpt-5`*	×	0.746	0.947	0.300	22.4 (412.8)	×	0.681	0.894	0.220	23.2 (605.6)	×	0.671	0.899	0.206	24.0 (676.8)
Category 2: w/ Syllable Constraint															
`claude-4.1-opus`	✓	0.691	0.907	0.376	12.2	✓	0.535	0.838	0.293	17.2	×	0.631	0.849	0.261	13.5
`deepseek-v3`	×	0.700	0.924	0.307	12.4	×	0.557	0.866	0.245	15.8	×	0.556	0.868	0.225	16.3
`gemini-2.5-pro`*	×	0.698	0.927	0.322	10.7 (1231.3)	×	0.517	0.853	0.266	12.3 (1450.1)	×	0.586	0.827	0.252	10.7 (554.5)
`gpt-5`*	×	0.705	0.928	0.345	21.7 (2024.9)	×	0.534	0.874	0.283	23.1 (1962.3)	×	0.586	0.833	0.270	21.0 (3105.8)
Category 3: Best of N															
`claude-4.1-opus`	✓	0.691	0.903	0.367	109.7	✓	0.611	0.848	0.324	109.2	✓	0.621	0.857	0.304	110.0
`deepseek-v3`	✓	0.691	0.902	0.332	122.6	✓	0.527	0.863	0.286	160.5	✓	0.522	0.851	0.275	113.6
`gemini-2.5-pro`*	✓	0.673	0.904	0.276	133.2 (1089.2)	✓	0.517	0.854	0.241	136.0 (972.7)	✓	0.554	0.858	0.235	133.8 (611.1)
`gpt-5`*	✓	0.673	0.906	0.296	114.2 (1810.2)	✓	0.571	0.861	0.275	111.2 (2229.6)	✓	0.529	0.855	0.256	110.8 (2158.8)
Category 4: Translate + Modify Pipeline (2 run)															
`claude-4.1-opus`	×	0.694	0.904	0.327	30.2	×	0.653	0.868	0.276	42.7	×	0.643	0.876	0.246	37.6
`deepseek-v3`	×	0.703	0.934	0.322	88.1	×	0.645	0.874	0.239	89.4	×	0.543	0.873	0.221	102.1
`gemini-2.5-pro`*	✓	0.691	0.902	0.348	22.3 (2517.7)	×	0.627	0.830	0.267	27.5 (1043.1)	✓	0.602	0.840	0.250	22.4 (2975.2)
`gpt-5`*	✓	0.690	0.917	0.321	42.2 (2666.2)	✓	0.514	0.842	0.275	46.6 (2773.0)	✓	0.557	0.820	0.234	44.9 (4012.9)
Category 5: HOMURA															
Ours (HOMURA$_{\text{Rubric}}$)	✓	0.700	0.914	0.378	9.2	✓	0.591	0.853	0.317	16.3	✓	0.620	0.863	0.260	13.1
Ours (HOMURA$_{\text{Reason}}$)	✓	0.701	0.925	0.376	10.1	✓	0.603	0.861	0.321	14.7	✓	0.605	0.858	0.309	13.2

Note: IB: In Bounds, ✓ indicates within bounds, × indicates out of bounds. AVG-Tokens: Average number of output tokens per sample; for reasoning models (*), values in parentheses denote the total tokens consumed (including internal thinking).

- **Category 2: Prompt-based Compression.** Category 1 models with brevity-oriented system prompts enforcing syllable-level conciseness.
- **Category 3: Best of N Refinement.** Generating N candidates with varying lengths, followed by a greedy selection of the highest-fidelity output that fulfills the syllable budget.
- **Category 4: Pipeline-based Post-editing.** Two-stage translation → length-constrained rewriting.
- **Category 5: Our Method (HOMURA).** RL fine-tuning on `Qwen3-8B` (Yang et al., 2025) to enforce temporal budgets as hard constraints.

More experimental details are provided in Appendix E.

4.2 Target Syllable Bounds Configuration

Following Section 2.3, we set density-calibrated target bounds $\mathcal{B}_L$ to impose comparable compression pressure across languages: En $\in [0.8, 0.9]$, De $\in [0.9, 1.0]$, and Es $\in [1.0, 1.1]$. These intervals are used as hard constraints for HOMURA.

4.2.1 Evaluation Metrics

We report complementary metrics for efficiency, semantic preservation, and linguistic quality:

Translation Density (BLEU-ρ ↑) We define **BLEU-ρ** as semantic fidelity per temporal unit:

$$\text{BLEU-}\rho = \frac{\text{BLEU}(x, \hat{x})}{\rho(x, y)}, \tag{14}$$

where $\text{BLEU}(x, \hat{x})$ is the back-translation BLEU (Papineni et al., 2002). Since BLEU alone may not fully capture fluency in back-translation paradigms (Edunov et al., 2020), we use BLEU-ρ specifically to measure **information density**, while linguistic quality assessed via Cometkiwi.

Semantic Integrity (BT-CERR ↑) BT-CERR is a binary indicator that evaluates whether all core events extracted from x are retained in the back-translation $\hat{x}$. (see Appendix D for the definition of core events).

Linguistic Quality (Cometkiwi ↑) We use reference-free Cometkiwi(Rei et al., 2022) to assess fluency and adequacy under compression.

4.3 Main Results

4.3.1 Cross-Lingual Performance and Temporal Compliance

Table 3 summarizes performance on Sand-Glass. We report human evaluation in Appendix C and Statistical Significance Testing in Appendix B. Overall, the results highlight the trade-off between temporal compliance and semantic density:

Budget Violation in Frontier Models. As shown by the IB column, unconstrained frontier model(CAT 1) universally fail to meet temporal budgets (×), particularly in low-density languages like Spanish. While specialized prompting (CAT 2) shifts the distribution toward brevity, it remains

unreliable, with models like `GPT-5` still exhibiting "verbosity drift" that exceeds target bounds. This confirms that pure instruction-following is insufficient for rigorous syllable-level constraints.

The Search-Fidelity Trade-off. Search-based (CAT 3) and pipeline (CAT 4) strategies achieve higher In-Bounds rates but at a significant semantic cost. For instance, CAT 4 variants often show a drop in BT-CERR, suggesting that the disjointed post-editing process leads to "hallucinated omissions" of critical source predicates. Moreover, the increased inference latency of Best-of-N makes it less practical in real-time setting.

Information Density Superiority of HOMURA. In contrast, HOMURA consistently maintains in bound while achieving the highest BLEU-ρ across all language pairs. This gap demonstrates that HOMURA does not merely truncate text but actively performs *semantic packing*—delivering a higher "semantic payload" per syllable. Additionally, we scale our method to a 32B model; see Appendix G for details.

Inference efficiency. HOMURA is markedly more inference-efficient. Unlike Best-of-N, which multiplies decoding cost by generating N candidates, and unlike translate-then-modify pipelines that require an extra rewriting pass, HOMURA satisfies the syllable/temporal budget in a single decoding. Consequently, it attains high IB compliance with substantially fewer tokens (AVG-Tokens in Table 3), making it suitable for real-time and high-throughput deployment.

Robustness of Reason-based GenRM. Within Category 5, $\text{HOMURA}_{\text{Reason}}$ achieves higher BT-CERR than $\text{HOMURA}_{\text{Rubric}}$ on two of the three target languages, with the largest gain observed for English. This suggests the reason-based reward better prunes redundancy while preserving the source's logical structure. We further validate this by replacing the self-trained GenRM in $\text{HOMURA}_{\text{Reason}}$ with an LLM-as-RM baseline (`DeepSeek-V3`); results are in Appendix J.

4.3.2 Analysis of Quality–Compression Trade-off

Figure 4 illustrates the trade-off between translation quality and output length. By plotting semantic metrics against the average syllable ratio ρ, we compare how different models perform as the available length budget decreases.

As shown in Figure 4, unconstrained baselines cluster in the high-quality, high-ρ region, representing the model's performance when length is not a factor. While standard LLMs can be prompted to shorten their outputs, their quality drops sharply as compression becomes more aggressive. This reflects a low rate–distortion efficiency: when forced to reduce the phonetic "rate" (syllables), these models suffer from disproportionately high "distortion" (loss of meaning), as they struggle to identify which information is most essential to preserve.

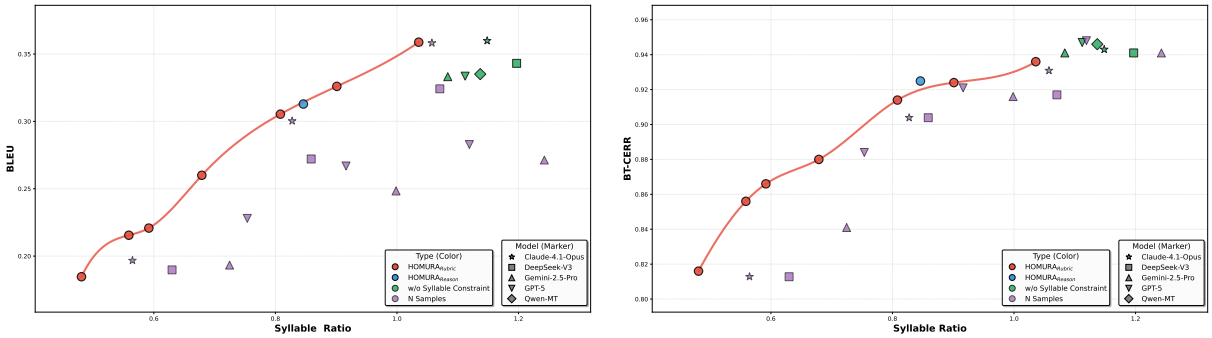
In contrast, HOMURA variants demonstrate significantly higher rate–distortion efficiency. At any given syllable budget, HOMURA maintains higher semantic fidelity than the baselines; conversely, it can achieve a much shorter length while maintaining the same quality level. These results suggest that our framework does not simply truncate text. Instead, it optimizes the model to operate at a more efficient trade-off point—packing a higher semantic payload into a limited phonetic budget to respect the strict constraints of media localization.

4.3.3 Qualitative Case Studies

We qualitatively compare HOMURA with representative baselines to assess practical behavior under different compression budgets. As shown in Table 10, frontier models (e.g., `GPT-5`, `Claude-4.1-opus`) often remain verbose even with brevity prompts, whereas HOMURA achieves tighter budget adherence by packing semantics into denser lexical and syntactic forms. When the budget becomes extreme (down to $\rho \approx 0.25$), HOMURA smoothly shifts from natural prose to highly condensed translations, prioritizing core predicate–argument content while discarding redundant function words. Full case comparisons and a taxonomy of the observed linguistic shifts are deferred to Appendix K.

4.4 Autonomous Compression: Navigating the Compression Limit

The preceding experiments show that HOMURA can reliably translate under a user-specified ratio interval. We further ask whether, without a preset lower bound, the model can *autonomously* discover the practical compression limit: the point where additional shortening triggers a sharp loss in meaning. To this end, we formulate an autonomous compression task by replacing the hard-interval length



(a) BLEU vs. Avg. Syllable Ratio

(b) BT-CERR vs. Avg. Syllable Ratio

Figure 4: Quality–compression trade-off on Zh→En. We plot BLEU (left) and BT-CERR (right) versus the average syllable ratio ρ. Each point is a model setting (marker: backbone; color: variant: Unconstrained, Rubric, GenRM, Multi-length Prompting). The red curve shows the empirical trend: quality degrades as compression strengthens.

Table 4: The Compression Wall. The autonomous model reliably reaches the feasible limit; stricter manual targets miss ρ and produce worse compression.

Model & Configuration	ρ	BLEU-ρ	BT-CERR
Auto HOMURA	**0.485**	0.404	0.811
Manually Constrained HOMURA			
$R_{\text{target}} \in [0.3, 0.4]$	0.558	0.392	0.819
$R_{\text{target}} \in [0.4, 0.5]$	**0.481**	0.384	0.816
$R_{\text{target}} \in [0.5, 0.6]$	0.559	0.386	0.856
$R_{\text{target}} \in [0.8, 0.9]$	0.808	0.378	0.914

reward with a continuous incentive:

$$R_{\text{len}}^{(\text{auto})}(x, y) = \min\left(\cos\left(\theta \cdot \rho(x, y)\right),\ 0\right), \quad (15)$$

where θ scales the incentive gradient. The policy is encouraged to minimize ρ while still being constrained by the quality reward R_{qual}.

The Empirical Compression Wall We observe a stable *compression wall* when pushing the model to its structural limit. In Table 4, autonomous optimization consistently converges to a narrow operating regime around $\rho \in [0.46, 0.49]$, regardless of the specific reward configuration.

When we enforce an ultra-aggressive target of $[0.3, 0.4]$, the model does *not* become shorter; instead it rebounds to $\rho = 0.558$ and semantic coherence collapses. This non-linear failure indicates a feasibility threshold around $\rho \approx 0.49$ for Zh→En: compressing beyond it creates an irreconcilable length-quality trade-off, where the model cannot preserve basic well-formedness while remaining brief.

Efficiency at the Frontier To quantify compression efficiency, we report BLEU-ρ. The autonomous model attains the best BLEU-ρ while operating at a genuinely shorter length ($\rho = 0.485$ vs. 0.558), indicating that autonomous optimization finds a more efficient operating point on the quality-compression frontier.

Rate-Distortion Efficiency at the Frontier Qualitatively, the autonomous model tends to compress by *semantic condensation*: dropping modifiers, selecting denser lexical items, and simplifying clause structures. In contrast, when forced toward an overly aggressive interval, outputs more often exhibit unstable shortening behaviors, reflecting the tension between further length reduction and preserving core meaning.

Takeaway: The Wall as a Feasibility Boundary Together, these results suggest an empirical feasibility boundary around $\rho \approx 0.49$ for time-constrained translation in our setting (Zh→En).We hypothesize that this boundary represents the minimal information carrier required to express core predicate-argument content.

5 Conclusion

We address systemic cross-lingual verbosity bias in LLM translation, a key obstacle in time-constrained settings such as subtitles and dubbing. We introduce Sand-Glass, a syllable-budgeted benchmark grounded in the Iso-Information Principle, and propose HOMURA, an RL framework that optimizes the fidelity–feasibility trade-off.

With a KL-regularized objective and a dynamic syllable-ratio reward, HOMURA achieves strict length compliance without reference translations while improving information density (BLEU-ρ). Future work will test whether the observed compression feasibility boundary (e.g., $\rho \approx 0.49$ for Zh→En) generalizes across language pairs, and extend HOMURA to end-to-end speech translation with joint semantic and duration optimization.

Limitations

First, while we adopt syllable count as a duration proxy grounded in the Iso-Information Principle, this metric remains a textual approximation of acoustic reality. In real-world dubbing and subtitling, physical duration is heavily influenced by prosodic features and speech tempo, and our current text-only framework does not account for multimodal constraints like lip-synchronization or isochrony. Consequently, generated outputs may still require minor timing adjustments during recording.

Second, our experimental validation is currently restricted to Chinese-to-English, German, and Spanish translation tasks. The universality of the "compression feasibility boundary" observed in our study remains to be verified across linguistically distant pairs or low-resource languages. Future work is needed to determine if such limits are language-specific or intrinsic to the translation task itself.

Acknowledgments

References

Anthropic. 2025. Claude 4.1 Opus System Card. Technical report, Anthropic.

Eleftheria Briakou, Zhongtao Liu, Colin Cherry, and Markus Freitag. 2024. On the implications of verbose llm outputs: A case study in translation evaluation. *arXiv preprint arXiv:2410.00863*.

Harveen Singh Chadha, Aswin Shanmugam Subramanian, Vikas Joshi, Shubham Bansal, Jian Xue, Rupeshkumar Mehta, and Jinyu Li. 2025. Length aware speech translation for video dubbing. *Preprint*, arXiv:2506.00740.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Christophe Coupé, Yoon Mi Oh, Dan Dediu, and François Pellegrino. 2019. Different languages, similar encoding efficiency: Comparable information rates across the human communicative niche. *Science Advances*, 5(9):eaaw2594.

L Davisson. 1972. Rate distortion theory: A mathematical basis for data compression. *IEEE Transactions on Communications*, 20(6):1202–1202.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Sergey Edunov, Myle Ott, Marc'Aurelio Ranzato, and Michael Auli. 2020. On the evaluation of machine translation systems trained with back-translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2836–2846, Online. Association for Computational Linguistics.

Marco Gaido, Sara Papi, Mauro Cettolo, Roldano Cattoni, Andrea Piergentili, Matteo Negri, and Luisa Bentivogli. 2024. Automatic subtitling and subtitle compression: Fbk at the iwslt 2024 subtitling track. In *Proceedings of the 21st International Conference on Spoken Language Translation (IWSLT 2024)*, pages 86–96.

Yuxuan Gu, Wenjie Wang, Xiaocheng Feng, Weihong Zhong, Kun Zhu, Lei Huang, Ting Liu, Bing Qin, and Tat-Seng Chua. 2025. Length controlled generation for black-box llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16878–16895.

Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. 2024. Llm-rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts. *arXiv preprint arXiv:2501.00274*.

Di He, Yingce Xia, Tao Qin, Laiwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc.

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. *arXiv preprint arXiv:2302.09210*.

Aditi Jha, Sam Havens, Jeremy Dohmann, Alex Trott, and Jacob Portes. 2023. Limit: Less is more for instruction tuning across evaluation paradigms. *arXiv preprint arXiv:2311.13133*.

Wenxiang Jiao, Wenxuan Wang, Jen-tse Huang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Is chatgpt a good translator? yes with gpt-4 as the engine. *arXiv preprint arXiv:2301.08745*.

Tollef Emil Jørgensen and Ole Jakob Mengshoel. 2025. Cross-lingual sentence compression for length-constrained subtitles in low-resource settings. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING 2025)*, pages 6447–6458.

Alina Karakanta, Matteo Negri, and Marco Turchi. 2020a. Is 42 the answer to everything in subtitling-oriented speech translation? In *Proceedings of the 17th International Conference on Spoken Language Translation*, pages 209–219, Online. Association for Computational Linguistics.

Alina Karakanta, Matteo Negri, and Marco Turchi. 2020b. MuST-cinema: A speech-to-subtitles corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3727–3734.

Wendi Li, Wei Wei, Kaihe Xu, Wenfeng Xie, Dangyang Chen, and Yu Cheng. 2024. Reinforcement learning with token-level feedback for controllable text generation. *arXiv preprint arXiv:2403.11558*.

Xiaobo Liang, Haoke Zhang, Juntao Li, Kehai Chen, Qiaoming Zhu, and Min Zhang. 2025. Generative reward modeling via synthetic criteria preference learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 26755–26769.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Dakota Mahan, Duy Van Phung, Rafael Rafailov, Chase Blagden, Nathan Lile, Louis Castricato, Jan-Philipp Fränken, Chelsea Finn, and Alon Albalak. 2024. Generative reward models. *arXiv preprint arXiv:2410.12832*.

Shushen Manakhimova, Vivien Macketanz, Eleftherios Avramidis, Ekaterina Lapshinova-Koltunski, Sergei Bagdasarov, and Sebastian Möller. 2024. Investigating the linguistic performance of large language models in machine translation. In *Proceedings of the Ninth Conference on Machine Translation*, pages 355–371.

OpenAI. 2025. GPT-5 System Card. Technical report, OpenAI.

Longshen Ou, Xichu Ma, Min-Yen Kan, and Ye Wang. 2023. Songs across borders: Singable and controllable neural lyric translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 447–467.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

François Pellegrino, Christophe Coupé, and Egidio Marsico. 2011. A cross-language perspective on speech information rate. *Language*, 87:539–558.

Tiago Pimentel, Clara Meister, Elizabeth Salesky, Simone Teufel, Damián Blasi, and Ryan Cotterell. 2021. A surprisal–duration trade-off across and within the world's languages. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 949–962, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ricardo Rei, Marcos Treviso, Nuno M Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José GC De Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, and 1 others. 2022. Cometkiwi: Ist-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645.

Le Ren, Xiangjian Zeng, Qingqiang Wu, and Ruoxuan Liang. 2025. Lyricar: A difficulty-aware curriculum reinforcement learning framework for controllable lyric translation. *arXiv preprint arXiv:2510.19967*.

Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. 2023. Verbosity bias in preference labeling by large language models. *arXiv preprint arXiv:2310.10076*.

Claude E Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *Preprint*, arXiv:2402.03300.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. 2023. A long way to go: Investigating length correlations in rlhf. *arXiv preprint arXiv:2310.03716*.

Joar Skalse, Nikolaus Howe, Dmitrii Krasheninnikov, and David Krueger. 2022. Defining and characterizing reward gaming. In *Advances in Neural Information Processing Systems*, volume 35, pages 9460–9471.

Yihan Wu, Junliang Guo, Xu Tan, Chen Zhang, Bohan Li, Ruihua Song, Lei He, Sheng Zhao, Arul Menezes, and Jiang Bian. 2023. Videodubber: machine translation with speech-aware length control for video dubbing. AAAI'23/IAAI'23/EAAI'23. AAAI Press.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *arXiv preprint arXiv:2506.05176*.

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.

A Related Work

A.1 Constrained Translation and Subtitling

Media localization imposes strict spatiotemporal constraints beyond semantic fidelity (Karakanta et al., 2020b). To address these limitations, early research focused on supervised learning paradigms that integrate specialized markers to guide line length and reading speed (Karakanta et al., 2020a). An advancement was introduced by Wu et al. (2023) with *VideoDubber*, which pioneered direct speech-length control by incorporating token-level duration predictions into the decoder stack. However, this approach remains constrained by its dependence on supervised training with non-isochronous datasets and heuristic duration labeling that introduces annotation noise. Building on this, Chadha et al. (2025) demonstrated that even with automatically generated length tags derived from phoneme ratios, supervised training on non-isochronous speech-translation data remains a fundamental bottleneck for achieving precise isochronicity. While architectures like CLSC (Jørgensen and Mengshoel, 2025) further refine this task using control tokens to manage sequence length, they remain fundamentally dependent on supervised training with task-specific compression corpora. In contrast, our work targets general-purpose Large Language Models (LLMs) via a reinforcement learning framework. This approach explicitly optimizes the trade-off between semantic preservation and temporal compliance without requiring ground-truth compression datasets.

A.2 Hard Constraints and Lyric Translation

The imposition of strict feasibility constraints parallels automatic lyric translation, where models must adhere to syllable counts and rhythm (Ou et al., 2023). Building on curriculum RL strategies for discrete constraints (Ren et al., 2025), *HOMURA* treats temporal duration as a hard constraint. We employ a KL-regularized objective to balance these rigid length requirements with semantic quality, ensuring the model respects the "sand-glass" temporal budget.

A.3 Verbosity Bias in LLMs

A major challenge for time-constrained translation is the systemic "verbosity bias" in LLMs, often exacerbated by RLHF alignment that conflates length with helpfulness (Saito et al., 2023; Skalse et al., 2022; Singhal et al., 2023). Our framework counteracts this by introducing a dynamic syllable-ratio reward. Instead of encouraging length drift, we specifically penalize verbosity, aligning the model's output with the strict density requirements of subtitling and dubbing.

B Statistical Significance Testing

We perform hypothesis testing to examine whether the improvements brought by HOMURA are statistically significant. For a fair comparison, for each baseline we choose its best-performing configuration that (i) requires a single inference pass and (ii) satisfies the target in-bounds constraint (IB), and compare it with HOMURA under the same constraints. Table 5 reports the resulting p-values.

Table 5: p-values for testing HOMURA against each variant.

Model	p-value of HOMURA					
	Zh → En			Zh → De		
	IB	BLEU-ρ	COMETKiwi	IB	BLEU-ρ	COMETKiwi
$\text{HOMURA}_{\text{Rubric}}$	✓	6.46e-03	8.20e-03	✓	$<1e-10$	$<1e-10$
$\text{HOMURA}_{\text{Reason}}$	✓	-	6.30e-03	✓	$<1e-10$	$<1e-10$

Across both language pairs and evaluation metrics, HOMURA yields statistically significant improvements over its strongest single-pass, in-bounds baselines, with p-values consistently below the conventional 0.01 threshold (often far smaller, e.g., $< 1e-10$ for Zh→De). These results confirm that the observed gains are unlikely to be due to random variation. The "–" entry denotes a degenerate case where the two compared results are identical, yielding no measurable difference and thus no meaningful significance test. For Zh→Es, no baseline satisfies the single-pass and in-bounds constraints, and thus we do not conduct hypothesis testing.

C Human Evaluation

We additionally report human evaluation results for the Zh-to-En setting to further validate the effectiveness of our method and the appropriateness of our evaluation metrics.

Table 6 shows that (i) without syllable constraints, all frontier LLMs achieve near-ceiling Accuracy/Completeness, serving as an unconstrained upper bound; (ii) enforcing syllable budgets by prompting substantially degrades human-rated quality, especially Accuracy; and (iii) Best-of-N provides limited and unstable gains. In con-

Table 6: Human evaluation results on Accuracy and Completeness under different settings. Note: IB: In Bounds, ✓ indicates within bounds, × indicates out of bounds.

Setting	Model	IB	Accuracy	Completeness	Avg.
w/o syll.	Gemini-2.5-Pro	×	4.95	5.00	4.97
	Claude-4.1-Opus	×	4.91	4.98	4.94
	GPT-5	×	4.96	5.00	4.98
	DeepSeek-V3	×	4.97	5.00	4.99
w/ syll.	Gemini-2.5-Pro	×	4.44	4.90	4.67
	Claude-4.1-Opus	✓	4.42	4.93	4.67
	GPT-5	×	4.42	4.87	4.65
	DeepSeek-V3	×	4.42	4.87	4.65
Best-of-N	Gemini-2.5-Pro	✓	4.57	4.57	4.57
	Claude-4.1-Opus	✓	4.51	4.49	4.50
	GPT-5	✓	4.71	4.80	4.75
	DeepSeek-V3	✓	4.56	4.58	4.57
Ours	$\text{HOMURA}_{\text{Rubric}}$	✓	4.91	4.83	4.87
	$\text{HOMURA}_{\text{Reason}}$	✓	4.85	4.93	4.89

trast, our HOMURA variants deliver the best overall constrained performance, approaching the unconstrained ceiling, with $\text{HOMURA}_{\text{Rubric}}$ slightly favoring Accuracy and $\text{HOMURA}_{\text{Reason}}$ favoring Completeness.

D The Sand-Glass Benchmark Construction

Sand-Glass is designed for *time-constrained* translation, where the output must satisfy a strict temporal budget while preserving essential meaning. The construction emphasizes two components: (i) **duration-aware segmentation** to derive realistic per-segment budgets from natural speech, and (ii) **core event extraction** to enable a fine-grained semantic retention check under forced compression (facilitating the *BT-CERR* metric defined in Sec. 4.2.1).

Pipeline Overview. Starting from real-world video datasets across diverse specialized domains, we extract ASR transcripts, segment them into duration-bounded subtitle units, apply multi-stage quality filtering (including strict length constraints), and finally extract *core events* (predicate–argument abstractions) to serve as a semantic backbone.

Construction Procedure.

1. **Data Acquisition and Segmentation.** We collect a large-scale corpus of real-world video transcripts covering five representative domains: *Gaming*, *Film & Television*, *Travel & Tourism*, *ACGN (Animation, Comics, Games and Novels)*, and *General Knowledge*. Speech is transcribed via ASR and segmented based on pauses and semantic units. Each segment is assigned a temporal budget derived from its actual speech duration.

2. **Preprocessing and Packaging.** We strip non-speech artifacts (e.g., fillers, music markers) and package segments into sliding windows of ten to preserve contextual coherence. Outputs are enforced into a fixed JSON schema for model interaction.

3. **Multi-stage Quality Filtering.** We apply a rigorous filtering pipeline to ensure structural and linguistic integrity. *Statistical Sanity:* To ensure the statistical stability of the ρ metric, we exclude segments with fewer than 10 characters or extreme Characters-Per-Second (CPS) values. This prevents short-sequence outliers from inflating density results and ensures that temporal constraints necessitate meaningful semantic compression. *Quality Heuristics:* Samples are scored via an ensemble of signals—including perplexity, repetition, and script consistency—retaining borderline cases with lower weights to preserve a natural difficulty curve. *Diversity:* Following LSH-based de-duplication (Jha et al., 2023), we perform *domain-balanced quota sampling* to obtain a high-quality final corpus of *1,000 golden instances*. This ensures a representative and manageable distribution of expansion challenges across the five selected domains.

4. **Core Event Extraction.** Since compression increases the risk of meaning-critical omissions, we extract *core events* from each source segment as a minimal semantic representation. Concretely, we identify the main predicate and collect its content-bearing arguments, including nouns, proper nouns, and numerals by using using Stanza Chinese POS tagging.[2]. We then measure semantic consistency by computing contextual embedding similarity between source and candidate predicate–argument realizations, and flag a violation if no candidate event sufficiently matches the source core event. This predicate–argument abstraction serves as the reference for measuring semantic integrity in compressed translations (via *BT-CERR*).

[2] https://stanfordnlp.github.io/stanza/

Table 7: Key hyperparameters used in training HOMURA.

Hyperparameter	Value
α_1	0.4
α_2	0.5
λ_{len}	0.5
λ_{qual}	0.5
k	300
τ_{min}	0
τ_{max}	0.8

E Implementation Details

Our RM training framework is built on Megatron. We use the Qwen3-32B-Chat model as the initialization. The training is conducted with a batch size of 256, using a cosine learning rate scheduler with an initial learning rate of 5e-6. All models are trained on 64 Huawei's Ascend 910B NPUs.

Our RL training framework is based on the Verl framework. We use the Qwen3-8B-Chat model as the initialization for RL training. During training, we configure a batch size of 16 and perform 16 rollouts per prompt using the GRPO algorithm. The learning rate is initialized at 1e-8, and a cosine scheduler with warm-up is applied toward the final iteration. Sampling is conducted with a temperature of 1.0, and the maximum generation length is limited to 1,024 tokens. The KL penalty coefficient β is set to 0, effectively removing the KL constraint relative to the reference policy. The PPO clipping range ϵ is fixed at 0.2. All models are trained for one epoch using 8 NVIDIA H800 80G GPUs. Other detail training hyperparameters are listed in Table 7.

Our training set consists of 70K instances drawn from the same data sources as the Sand-Glass benchmark, and is constructed following the same procedure.

F Effectiveness of Dynamic Syllable-ratio Reward

In Section 3.2.1, we hypothesized that rigid syllable-ratio bounds are susceptible to discretization noise, particularly in shorter utterances. To validate the necessity of our dynamic relaxation mechanism, we conduct an ablation study comparing our *Dynamic Syllable-ratio Reward* against a static baseline on the Zh→En translation task. Both methods share the same target objective ($\mathcal{B}_L \in [0.8, 0.9]$), but the static version enforces fixed boundaries regardless of source length.

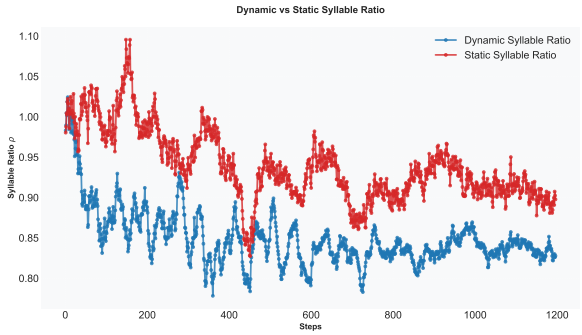


Figure 5: Training dynamics of syllable ratios during policy optimization (Zh→En). The blue curve represents our dynamic bound relaxation, while the red curve represents static fixed bounds. Values indicate the mean syllable ratio ρ across training steps.

As shown in Figure 5, the experimental results yield several key observations:

- **Stability and Convergence:** The Dynamic reward (blue) demonstrates significantly smoother optimization and faster convergence into the target interval $[0.8, 0.9]$. In contrast, the Static reward (red) exhibits violent oscillations, especially in early training phases.
- **Mitigating Discretization Noise:** The instability of the static baseline confirms that "hard" rewards for syllable counts are overly sensitive to small variations in short segments. A single-syllable difference in a short sentence can cause a disproportionate jump in the ratio, providing an inconsistent gradient for the RL agent.
- **Optimization Efficiency:** By adaptively relaxing the lower bound for shorter sequences via $\gamma(x)$, our method provides a more continuous and achievable reward landscape. This prevents the model from being penalized for unavoidable linguistic granularity, leading to more efficient learning of the desired compression behavior.

These findings justify the use of length-aware dynamic bounds in HOMURA to ensure stable policy refinement across diverse utterance lengths.

G Scaling to Larger Model.

We analyze the effect of model scaling by comparing 8B and 32B variants of HOMURA under identical training settings, with results reported in Table

Table 8: Scaling Analysis of HOMURA (Zh→En). Comparison between 8B and 32B backbones at $\mathcal{B}_L \in [0.8, 0.9]$.

Scale	Model	IB	Cometkiwi	BT-CERR	BLEU-ρ
8B	**HOMURA**$_{\text{Rubric}}$	✓	0.700	0.914	0.378
	HOMURA$_{\text{Reason}}$	✓	0.701	0.925	0.376
32B	**HOMURA**$_{\text{Rubric}}$	✓	0.740	0.917	0.388
	HOMURA$_{\text{Reason}}$	✓	0.740	0.933	0.385

8. Scaling consistently improves translation quality and semantic preservation, as shown by higher BLEU-ρ, Cometkiwi and BT-CERR scores. Moreover, the relative advantage of HOMURA$_{\text{Reason}}$ over HOMURA$_{\text{Rubric}}$ is preserved at larger scale, suggesting that the proposed reward design scales robustly with model capacity.

H Examples of Structured CoT Reasoning for GenRM

This appendix provides illustrative examples of the structured CoT reasoning produced by GenRM.

Structured CoT Example #1

Source sentence (x):
我曾经在这看到过几艘陌生的船。

Translation hypothesis (y):
I have seen a few strange ships here before.

COT process:

1. **Back-translation:** The translation "I have seen a few strange ships here before." is back-translated into Chinese. The back-translation result is "我以前曾在这里见过几艘奇怪的船。" (I have previously seen several strange ships here). Alternatively, it can also be "我曾经在这里看到过几艘陌生的船。" (I once saw several unfamiliar ships here).

2. **Semantic consistency judgment:** Compare the source sentence "我曾经在这看到过几艘陌生的船。" (I once saw several unfamiliar ships here) with the back-translated content "我以前曾在这里见过几艘奇怪的船。" (I have previously seen several strange ships here). The two are completely consistent in core semantics, both expressing the fact of "过去在某地看到过不熟悉的船只" (having seen unfamiliar ships at a certain place in the past). "曾经" (once / before) corresponds to "before," "在这" (here) corresponds to "here," "看到过" (have seen) corresponds to "have seen," "几艘" (a few) corresponds to "a few," "陌生的" (unfamiliar / strange) corresponds to "strange," and "船" (ships) corresponds to "ships." The semantic consistency is very high.

3. **Translation quality judgment:** The translation "I have seen a few strange ships here before." is grammatically correct, using the present perfect tense to express a past experience, which conforms to English usage. The expression is fluent and natural, with accurate word choice. Therefore, the translation quality is high.

Based on the above, the semantics are consistent and the translation quality is high, resulting in a final score of 1.

Score:
1

Structured CoT Example #2

Source sentence (x):
所以天命人只有超度了金池长老的冤魂

Translation hypothesis (y):
So the Mandate Person can only rescue the lingering soul of Jinchi Elder

COT process:

1. **Back-translation:** The translation "So the Mandate Person can only rescue the lingering soul of Jinchi Elder" is back-translated into Chinese as "所以天命人只能解救金池长老逗留的灵魂。" (So the Mandate Person can only rescue the lingering soul of Elder Jinchi).

2. **Semantic consistency judgment:** The key terms in the source text are "超度" (perform salvation rites / help a soul attain deliverance) and "冤魂" (wronged or vengeful spirit). "超度" is a term with strong religious and cultural connotations, referring to helping the deceased soul attain liberation and enter reincarnation or the afterlife through rituals. However, the word "rescue" in the translation is overly

general and fails to convey the deeper religious meaning of "超度". "冤魂" refers to a ghost that died unjustly and carries strong resentment, whereas "lingering soul" (逗留的灵魂) only expresses the idea of lingering or wandering, losing the core semantic component of "冤" (wronged / resentful). Therefore, the back-translated content is semantically inconsistent with the original text.

3. **Translation quality judgment:** Due to the semantic inconsistency, this step is skipped.

Based on the above analysis, the semantics are inconsistent, and the final score is 0.

Score:
0

I Prompt for Linguistic well-formedness Reward.

This appendix provides the training prompt used for the linguistic well-formedness reward.

Training Prompt
You are an expert in {target_lang} translation. Given context `<context>` for background reference. Here is a text `<text>` from the context and its corresponding {target_lang} translation `<translation>`. You need to determine whether `<translation>` is a qualified {target_lang} translation.

Requirements: Consider only the following criteria; if any are not met, the score is 0:

1. **Grammaticality:** No obvious grammatical or syntactic errors, consistency errors (e.g., subject–verb agreement, tense, case, gender/number), collocation errors, or linguistic defects caused by punctuation or spelling.

2. **Readability & Coherence:** The sentence structure is complete, logically coherent, and naturally connected, making it smooth and easy to understand; there should be no obvious fragmentation, garbled text, repetitive stacking, or malformed sentences.

Input:
`<context>`: {context}
`<text>`: {text}
`<translation>`: {translation}

Output Requirements:

- If `<translation>` satisfies both criteria above, return «`1`».
- Otherwise, return «`0`».
- Please output «`0`» or «`1`» directly; do not output any explanation.

Output:

J Ablation of GenRM Design

This appendix presents an ablation study on the design of the reward model within the GenRM framework. In particular, we compare our self-trained GenRM with an external LLM-as-RM. In this ablation, the external reward model fully replaces GenRM, while all other components of the compression-oriented reinforcement learning pipeline are kept identical.

External LLM-as-RM Setup For the external reward model baseline, we adopt an LLM-as-RM setup using `DeepSeek-V3`. Given a source sentence and its translation, the model is prompted to perform a quality assessment. Specifically, it evaluates whether the translation (i) preserves the core semantic content of the source sentence under the target syllable constraint, and (ii) remains fluent and grammatical. The model outputs a binary acceptability judgment, which is directly used as the reward signal during policy optimization. The prompt for external LLM-as-RM is below:

Prompt for GenRM

You are a translation quality reward model (GenRM).
Given contextual information, a source sentence, and its translation, you will perform step-by-step reasoning to assess translation quality. Your evaluation process must strictly follow the sequence below.
Step 1: Back-Translation
Translate the given translation back into the source language.

Step 2: Semantic Consistency Assessment
Compare the semantic consistency between the original source text and the back-translated text.

- If the meanings are consistent or highly consistent, proceed to Step 3.
- If they are inconsistent, skip Step 3 and assign a final score of 0.

Step 3: Translation Quality Assessment
Evaluate whether the translation is of high quality based on factors such as fluency, grammatical correctness, and cultural appropriateness.

Your final output must include a clear chain-of-thought reasoning process and a reward score:

- If semantic consistency is satisfied and the translation quality is high, set `score = 1`.
- Otherwise, set `score = 0`.

Output Format
The output must strictly follow the format below. Do not add or omit any fields:

```
{
"COT": "<COT reasoning>",
"score": 0 or 1
}
```

Input Format
Context: `{context}`
Source text: `{current_text}`
Translation: `{translated_text}`

Results and Analysis Table 9 compares compression translation models optimized with different reward model designs. The model trained with the self-trained GenRM consistently outperforms its counterpart using an external LLM-as-RM baseline, demonstrating that a reward model trained specifically for compression-oriented translation provides more effective optimization guidance than a generic, prompt-based evaluator.

K Case Study

This appendix provides a fine-grained analysis of the linguistic strategies employed by HOMURA to navigate the quality-compression frontier. Unlike standard LLMs that often rely on simple truncation, our model learns a hierarchy of compression operations.

Table 9: Effect of GenRM choice on compressed translation under $\text{HOMURA}_{\text{Reason}}$ (Average across all language pairs).

Model Variant	Cometkiwi	BT-CERR	BLEU-ρ
$\text{HOMURA}_{\text{Reason}}$	0.636	0.881	0.335
$\text{HOMURA}_{\text{Reason}}$ (w/ External RM)	0.628	0.879	0.330

Observation 1: Taxonomic Shifts in Semantic Packing. Analysis of Table 10 reveals three distinct tiers of semantic packing performed by HOMURA as constraints tighten:

- **Lexical Consolidation**: Mapping verbose multi-word phrases to high-density synonyms (e.g., the 7-syllable "relationship between them" $\rightarrow$ 1-syllable "bond").
- **Aspectual Simplification**: Reducing periphrastic verbal constructions into synthetic forms (e.g., "is becoming closer" $\rightarrow$ "grows closer"), which preserves the temporal aspect while reducing syllable count.
- **Syntactic Pruning**: Selectively removing low-surprisal functional tokens (articles, auxiliaries) while anchoring the sentence around core predicate-argument structures.

Observation 2: Behavioral Transition at the Compression Wall. The trade-off spectrum in Part 2 of Table 10 visualizes the transition from naturalistic translation to "telegraphic speech." While fluency is maintained down to $\rho \approx 0.50$, pushing toward the empirical limit ($\rho \approx 0.25$) forces the model to prioritize *propositional content* over *morphosyntactic correctness*. This behavior confirms that HOMURA does not randomly drop tokens but strategically re-allocates the syllable budget to the most informative constituents (e.g., "Bond grows tighter"), ensuring minimal meaning loss even at the threshold of reward collapse.

L Detailed Breakdown of Expansion Metrics

To provide a comprehensive view of the verbosity bias identified in Section 2, we present a fine-grained statistical breakdown of expansion metrics across various frontier LLMs and language pairs. As summarized in Table 11, the *Roundtrip Expansion Ratio* (R_{rtp}) consistently exceeds the unity baseline across all tested models and languages,

Table 10: Qualitative Case Study on Zh → En: Comparison between frontier LLMs and HOMURA across varying compression intensities. Source: 两人之间的关系越来越亲密 (12 syllables).

Category	Model	Output Text	Syl.	ρ	Quality Observation
Part 1: Comparison with Baselines (Target $\rho \in [0.8, 0.9]$ for Prompted Models)					
Unconstrained	Claude-4.1-Opus	The relationship between the two is becoming increasingly intimate.	20	1.67	Naturally verbose; high fidelity.
	Gemini-2.5-Pro	The relationship between the two grew increasingly intimate.	17	1.42	Fluent but ignores brevity.
	GPT-5	The relationship between the two is growing increasingly close.	17	1.42	Optimal quality; no length control.
Prompt-based Compression	Claude-4.1-Opus	Their relationship grows increasingly intimate.	13	1.08	Fails to reach target interval.
	Gemini-2.5-Pro	Their relationship is growing more intimate.	12	1.00	Fails to reach target interval.
	GPT-5	The two are becoming more and more intimate.	12	1.00	Fails to reach target interval.
Part 2: Controllable Compression via HOMURA *(ρ Trade-off Analysis)*					
HOMURA	$\rho \approx 1.25$	The relationship between the two is becoming closer.	15	1.25	Full fidelity and naturalness.
	$\rho \approx 0.83$	The bond between them is growing closer.	10	0.83	**Precise adherence** via condensation.
	$\rho \approx 0.58$	Bond between them grows closer.	7	0.58	Minor grammatical omission.
	$\rho \approx 0.50$	Their bond grows closer.	6	0.50	Extreme brevity; core meaning only.
	$\rho \approx 0.25$	Bond grows tighter.	3	0.25	**Compression Limit**; loss of fluency.

with a high percentage of segments ($> 60\%$) exhibiting systemic inflation.

The distribution shifts visualized in Figure 6 further confirm that this bias is not merely a result of specific difficult cases but represents a structural characteristic of current LLM-based translation.

M Declaration of AI Assistance

To enhance development efficiency, GitHub Copilot was utilized as a coding assistant in this study. Specifically, it facilitated the implementation of data preprocessing modules, the execution of statistical analyses, and the rendering of complex experimental figures. The core algorithmic design, reinforcement learning frameworks, and the interpretation of results remain the original work of the authors, who maintain full responsibility for the code's accuracy.

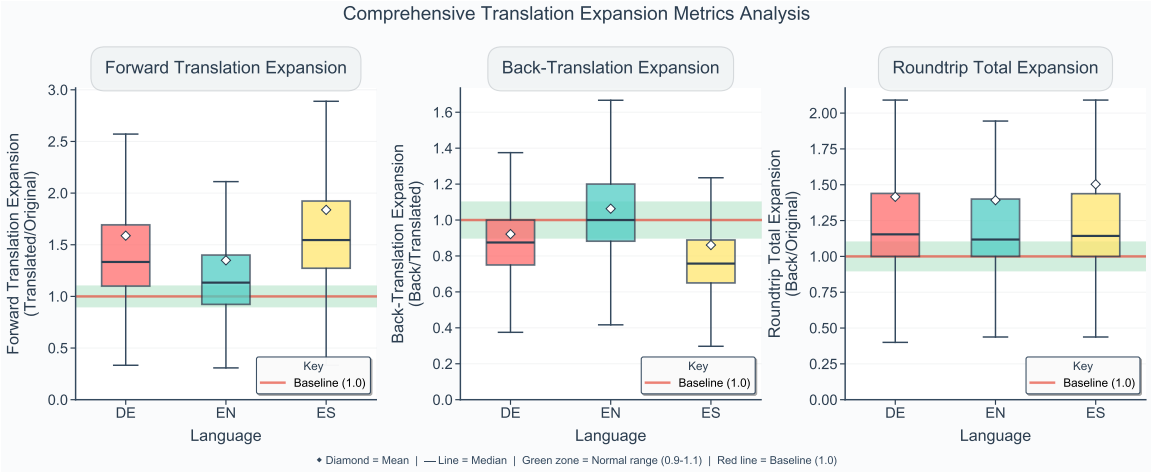


Figure 6: Statistical distribution of translation expansion metrics across DE, EN, and ES. The boxplots illustrate the variations in Forward, Back, and Roundtrip expansion ratios, with diamonds indicating means.

Table 11: Detailed Breakdown of Expansion Metrics across Models and Language Pairs. R_{fwd}, R_{bwd}, and R_{rtp} denote Forward, Backward (Back-translation), and Roundtrip Expansion Ratios, respectively. Values are presented as Mean and Standard Deviation.

Language	**Model**	R_{fwd} **(Mean ± Std)**	R_{bwd} **(Mean ± Std)**	R_{rtp} **(Mean ± Std)**	$R_{rtp} > 1$ **(%)**
De	`claude-4.1-opus`	1.58 ± 0.96	0.95 ± 1.54	1.39 ± 0.99	66.2
	`deepseek-v3`	1.67 ± 1.37	0.91 ± 0.29	1.54 ± 1.89	63.8
	`gemini-2.5-pro`	1.51 ± 0.88	0.96 ± 0.44	1.35 ± 0.84	59.0
	`gpt-5`	1.63 ± 0.96	0.90 ± 0.19	1.43 ± 0.87	71.8
En	`claude-4.1-opus`	1.36 ± 1.13	1.07 ± 0.55	1.41 ± 1.32	64.8
	`deepseek-v3`	1.41 ± 1.15	1.05 ± 0.48	1.47 ± 1.71	59.1
	`gemini-2.5-pro`	1.29 ± 0.81	1.07 ± 0.25	1.33 ± 0.80	61.0
	`gpt-5`	1.33 ± 0.81	1.07 ± 0.23	1.36 ± 0.80	65.9
Es	`claude-4.1-opus`	1.81 ± 1.05	0.90 ± 2.48	1.41 ± 1.19	66.9
	`deepseek-v3`	1.99 ± 1.83	1.01 ± 1.12	1.99 ± 2.90	64.0
	`gemini-2.5-pro`	1.72 ± 0.99	0.83 ± 0.36	1.35 ± 0.87	60.4
	`gpt-5`	1.86 ± 1.09	0.78 ± 0.18	1.40 ± 0.83	68.5