Reasoning as Supportive Context for Machine Translation: A Case Study on Hindi to Bengali Language Pair

Kshetrimayum Boynao Singh[1,2]*, **Saksham Singh**[1]*, **Partha Pakray**[2], **Asif Ekbal**[1]
Department of Computer Science and Engineering
[1]Indian Institute of Technology Patna, India
[2]National Institute of Technology Silchar, India
{boynfrancis, meetsakshamsingh, pakraypartha, asif.ekbal}@gmail.com

Abstract

We investigate whether reasoning information can enhance machine translation when incorporated as supportive context during training and inference. Using Hindi-Bengali translation as a case study, we define five reasoning components: Key Terms, Syntactic, Semantic, Pragmatic, and Paraphrase. We conduct a complete ablation across all 31 possible combinations using Gemma-3-1B-Instruct and evaluate on multi-domain benchmark with BLEU, chrF, and TER. Evaluation results show that reasoning effectiveness depends on its type and composition rather than quantity. Combining multiple heterogeneous signals causes objective diffusion, degrading performance. The compact Semantic and Paraphrase combination proves optimal, and providing it during inference yields 23.86 BLEU compared to 22.12 from standard fine-tuning a +1.74 BLEU gain across eight domains. These findings demonstrate that targeted semantic guidance consistently and meaningfully improves the compact translation models.

1 Introduction

Neural machine translation (MT) has advanced substantially with multilingual pretraining and instruction-tuned language models. However, translation quality remains uneven across many language pairs, especially when model capacity is limited. This is true even for related Indic languages, such as Hindi and Bengali, where lexical overlap coexists with differences in script, morphology, register, and domain-specific usage.

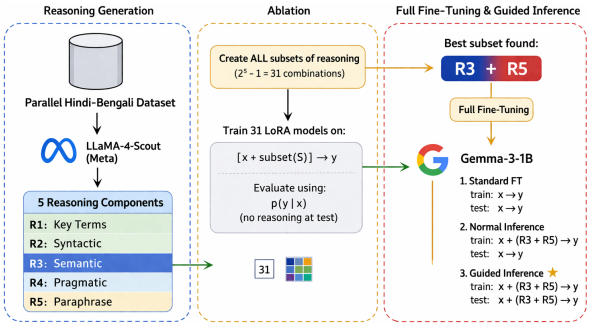

Figure 1: Overview of our reasoning-augmented MT methodology for Hindi-Bengali translation

Recent work has shown that reasoning signals can improve performance on tasks, such as arithmetic, logical inference, and commonsense reasoning (He et al., 2020). These findings have motivated the use of reasoning-oriented supervision in a broader range of natural language processing (NLP) problems. However, most prior work studies settings in which intermediate reasoning is either directly useful for the final prediction or is explicitly provided during inference (Lee et al., 2025) (Ghazaryan et al., 2025). Machine translation differs in an important way: the goal is not to generate reasoning traces, but to produce a fluent target sentence that faithfully preserves the meaning of the source. This raises a natural question: can reasoning information improve machine translation when it is used only as supportive context ? On the one hand, such information may clarify sentence meaning and foreground semantically important content. On the other hand, it may increase input complexity, introduce noisy supervision, or create a difference between training and inference conditions.

*Equal contribution.

To study this, we investigate reasoning as supportive context for Hindi–Bengali translation. During fine-tuning, the model learns from structured reasoning annotations paired with parallel data, and at inference, it uses both source text and reasoning context to generate translations. We define five components: *Key Terms*, *Syntactic*, *Semantic*, *Pragmatic*, and *Paraphrase*, and perform ablations over all non-empty combinations, yielding 31 reasoning-based models alongside a standard baseline. All models are fine-tuned from Gemma-3-1B-Instruct and evaluated using BLEU, chrF, and TER (Papineni et al., 2002).

Results show that reasoning impact varies across setups; adding more components does not ensure better performance. Effectiveness depends on composition, not quantity. A compact *Semantic+Paraphrase* combination performs best and remains stable, indicating targeted reasoning is beneficial.

2 Related Work

2.1 Reasoning in Large Language Models

Recent advances in large language models (LLMs)(Zhu et al., 2024; He et al., 2024) have sparked growing interest in reasoning-oriented (Nguyen and Xu, 2025) supervision. Studies on chain-of-thought (Wang et al., 2025; Jiang et al., 2025) prompting, tree of thoughts (Yao et al., 2023), and related reasoning strategies (Fan et al., 2025), such as self-consistency and least-to-most prompting show that models can achieve significant gains on tasks, such as arithmetic reasoning, commonsense inference (Liu et al., 2023), and symbolic problem (Gaur and Saunshi, 2023) solving. In these settings, reasoning steps help models structure their internal representations (Dabre et al., 2023) and guide (Jiang et al., 2025) the generation of correct outputs. However, most of this work focuses on tasks where reasoning is directly part of the solution process. Machine translation differs from these scenarios because the final output is a fluent sentence rather than an explicit reasoning trace, making it unclear whether reasoning supervision transfers effectively to translation tasks.

2.2 Machine Translation for Indic Languages

Neural machine translation for the Indic languages has improved through multilingual pretrained models (Bala Das et al., 2023; Singh et al., 2025a; Singh et al., 2026a), larger parallel datasets and dedicated benchmarking efforts. These systems have enhanced cross-lingual transfer (Singh et al., 2023) and improved translation quality across several South Asian languages. Nevertheless, many approaches rely on large-scale models and multilingual training pipelines (Verma et al., 2025). Closely related language pairs, such as Hindi and Bengali, still pose meaningful challenges due to differences in script, morphological structure, and domain-specific language use. As a result, translation accuracy depends not only on lexical similarity but also on reliable semantic interpretation.

2.3 Translation Support Beyond Parallel Data

Several studies have investigated enriching translation models with additional forms of guidance beyond plain parallel data. These approaches include terminology constraints, syntactic annotations, alignment information (Chen et al., 2017) domain labels (Yan et al., 2018; Singh et al., 2026b), and paraphrastic variants. Such signals aim to help models preserve technical terminology (Gao et al., 2025), resolve structural ambiguity and capture alternative expressions of the same meaning. However, combining multiple support signals reflects nuanced interactions rather than cumulative improvements (Singh et al., 2025b). Auxiliary information may overlap in function and influence the model's attention, shaping how effectively the central translation objective is emphasized.

2.4 Research Gaps in Reasoning for MT

Despite growing interest in reasoning supervision, its role in MT remains underexplored. Existing work rarely examines whether reasoning-style information can improve translation when used during training time, as well as inference time, and how the translation empowers. Furthermore, examining which path of reasoning signals are beneficial for translation and how different forms of the other supportive information interact when combined. This work addresses these gaps by decomposing reasoning into multiple linguistic components and conducting a complete combinational ablation study. By evaluating all the reasoning subsets, we identify which forms of reasoning supports and improve the translation quality, and which combinations degrade it, as well as which reasoning section is required during the inference time in order to improve translation capabilities.

3 Dataset

3.1 Parallel Corpus

We conduct experiments on a Hindi–Bengali parallel corpus spanning eight domains, *viz.* Agriculture, Tourism, Governance, Climate, Healthcare, Science and Technology, Judiciary, and Education. After filtering for sentence quality and annotation completeness, the final training set contains 36,040 sentence pairs, and the test set contains 2,000 sentence pairs. The test set supports both overall and domain-wise evaluation, with 200 examples for each domain except Education, which contains 600 examples.

The training corpus contains 771,621 Hindi tokens and 603,131 Bengali tokens, with average sentence lengths of 21.4 and 16.7 tokens, respectively. Bengali also shows higher lexical diversity, with 61,284 unique word types compared to 45,170 in Hindi. The test set follows a similar pattern, containing 41,027 Hindi tokens and 32,061 Bengali tokens. Overall, the corpus provides a diverse benchmark for evaluating translation quality across domains with different linguistic characteristics, ranging from technical and descriptive text to formal institutional and instructional language.

3.2 Data Filtering and Quality Considerations

Before training, we filter the corpus to remove noisy or misaligned sentence pairs. This process enforces basic quality constraints, such as reasonable sentence length, alignment consistency, and valid reasoning annotations. Examples with incomplete, empty, or malformed reasoning fields are excluded. These steps reduce training noise and improve the reliability of the subsequent ablation analysis.

3.3 Reasoning Annotation Pipeline

We generate structured reasoning annotations for each training sentence using Llama-4-Scout-17B-16E-Instruct. To leverage its stronger reasoning ability in English, each Hindi–Bengali sentence pair is first mapped through an intermediate Hindi–English translation, from which reasoning annotations are produced and then adapted back into Hindi and Bengali. To ensure scalability and consistency, the reasoning context is systematically transferred using Google Translate. The resulting annotations are designed to provide supportive contextual signals for translation rather than exhaustive linguistic analysis. For each sentence pair, the pipeline generates five reasoning components: Key terms (R_1), Syntactic (R_2), Semantic (R_3), Pragmatics (R_4), and Paraphrase (R_5). When applicable, annotations are produced for both Hindi and Bengali to maintain cross-lingual alignment. Structurally inconsistent annotations are filtered before training, yielding stable and uniform inputs for all 31 reasoning-based ablation configurations.

3.4 Reasoning Components

The five reasoning components capture complementary types of translation-relevant information.

R_1: **Key terms** identify important content words, explain their role in context, and clarify ambiguous expressions.

R_2: **Syntactic** information describes grammatical structure, including dependencies, modifiers, and clause boundaries, to support preservation of compositional meaning.

R_3: **Semantic** reasoning captures the core meaning of a sentence by identifying entities, actions, and their relations.

R_4: **Pragmatics** models discourse-level cues, such as speaker intent, tone, and contextual dependence beyond literal sentence meaning.

R_5: **Paraphrase** provides semantically equivalent reformulations with different lexical or syntactic realizations while preserving the original meaning.

These components differ not only in function but also in length. As shown in Table 1, Syntactic and Key Terms are the most verbose annotations, averaging more than 80 tokens per instance, whereas Semantic and Paraphrase are substantially more compact, averaging 45 and 24 tokens, respectively. This variation allows us to study both the effect of reasoning type and the effect of reasoning scale on translation performance.

4 Methodology

Our goal is to examine whether structured reasoning can serve as supportive context for machine translation, and how its presence or absence at inference time affects performance. To this end, we adopt a three-phase methodology: reasoning augmentation using *Llama-4-Scout-17B-16E-Instruct*, an exhaustive ablation study on *google/gemma-3-1b-it*, and an evaluation of inference settings to determine which reasoning configuration best improves translation quality.

Training data	Sentences	Tokens	Vocabulary	# Characters	Avg Tokens/Sent
Hindi (S)	36,040	771,621	45,170	4,008,578	21.4
Bengali (T)	36,040	603,131	61,284	3,930,184	16.7
Hindi Key-terms (R_1)	36,040	2,946,267	48,260	15,414,987	81.7
Bengali Key-terms (R_1)	36,040	2,226,749	90,354	16,424,124	61.8
Hindi Syntactic (R_2)	36,040	3,077,551	66,757	14,149,631	85.4
Bengali Syntactic (R_2)	36,040	2,487,377	110,260	8,353,154	69.0
Hindi Semantic (R_3)	36,040	1,628,938	28,499	4,506,294	45.2
Bengali Semantic (R_3)	36,040	1,194,331	48,063	14,561,377	33.1
Hindi Pragmatics (R_4)	36,040	2,675,864	32,384	16,420,421	74.2
Bengali Pragmatics (R_4)	36,040	2,001,633	49,014	7,986,367	55.5
Hindi Paraphrase (R_5)	36,040	857,822	24,545	13,794,926	23.8
Bengali Paraphrase (R_5)	36,040	633,695	43,157	4,342,617	17.6

Table 1: Training corpus statistics for the Hindi-Bengali parallel data and the five reasoning components on both sides

Domain (Sentence)	Language	Tokens
Agriculture (200)	Hindi	3,956
	Bengali	3,249
Tourism (200)	Hindi	4,028
	Bengali	3,222
Governance (200)	Hindi	4,160
	Bengali	3,319
Climate (200)	Hindi	4,093
	Bengali	3,296
Healthcare (200)	Hindi	4,049
	Bengali	3,174
Science_Technology (200)	Hindi	4,117
	Bengali	3,218
Judiciary (200)	Hindi	4,191
	Bengali	3,112
Education (600)	Hindi	12,433
	Bengali	9,471
Overall (2000)	Hindi	41,027
	Bengali	32,061

Table 2: Domain-wise and overall statistics of the Hindi-Bengali test set, including sentence counts and token counts for Hindi and Bengali

4.1 Problem Formulation

Let x denote a Hindi source sentence and y its Bengali translation. In standard machine translation, the model learns the conditional mapping $p(y \mid x)$. We extend this setting by augmenting the input with structured reasoning. Let $\mathcal{R} = \{R_1, R_2, R_3, R_4, R_5\}$ denote the five reasoning components. For any subset $S \subseteq \mathcal{R}$, the model is trained on the inputs formed by concatenating x with the reasoning components in S, while the target remains the translation y.

We evaluate under two inference paradigms: **Zero-Reasoning Inference:** Evaluation under $p(y \mid x)$, where reasoning is not provided at test time. This isolates the effect of training-time exposure to reasoning.

Reasoning Guided Inference: Evaluation under $p(y \mid x, S_{\text{optimal}})$, where the best reasoning subset is identified during ablation study and is provided along with the source during testing.

4.2 Base Model and Reasoning Pipeline

We use *Gemma-3-1B-Instruct* as the translation backbone, as it is compact enough for auxiliary context to meaningfully affect translation behavior. All reasoning components (R_1 to R_5) for both training and test data are generated using *Llama-4-Scout-17B-16E-Instruct*. Reasoning is first produced in English and then translated into Hindi and Bengali to maintain structural consistency. To avoid target leakage, test-time reasoning in Guided Inference is generated strictly from the Hindi source sentence, without access to the gold Bengali translation.

4.3 Structured Training Format

Each training instance is formatted as a prompt containing the Hindi source sentence, the selected reasoning components in Hindi and Bengali, and an instruction to generate only the Bengali translation. The loss is applied only to the translation output, ensuring that the model is not trained to reproduce the reasoning itself. To identify which reasoning signals are useful, we perform a complete ablation over all non-empty combinations of the five components using LoRA. This results in $2^5 - 1 = 31$ reasoning-conditioned models, along with a standard source-target finetuning. All 31 models are trained under identical settings: 1 epoch, learning rate 1×10^{-4} with cosine scheduling, batch size 4, and gradient accumulation over 16 steps, yielding an effective batch size of 64. We use `paged_adamw_8bit` optimizer (Kingma and Ba, 2017) in `bfloat16` precision. During this phase, evaluation is conducted using Zero-Reasoning Inference to identify the optimal subset S_{optimal}.

4.4 Full Parameter Fine-Tuning and Inference Evaluation

After identifying the optimal combination ($R_3 + R_5$), we perform full parameter fine-tuning to evaluate the effect of training-inference alignment more rigorously. We consider three setups:

1. **Standard Finetune:** Full fine-tuning on

source-target pairs only, evaluated under zero-reasoning inference.

2. **Normal Inference:** Full fine-tuning with $R_3 + R_5$ during training, but evaluation without reasoning inference. This measures the penalty caused by reasoning when is not provided at during the inference time.

3. **Guided Inference:** Full fine-tuning with $R_3 + R_5$ during training, and evaluation with the same $R_3 + R_5$ context at test time. This removes the difference between training and testing and measures the full benefit of reasoning-augmented translation.

5 Results

Our evaluation proceeds in two stages. First, we conduct an exhaustive *LoRA* ablation study to assess the effect of reasoning complexity and identify the most effective supportive context. As a reference point, the original baseline *Gemma-3-1B-Instruct* model without task-specific fine-tuning achieves 4.99 BLEU, 28.87 chrF, and 137.41 TER on the test set, underscoring the need for task-specific adaptation. Second, we evaluate full-parameter fine-tuning to examine how performance changes when reasoning is not provided at test time versus when it is. The complete results for all 31 reasoning configurations are presented in Table 3.

Group	Model	BLEU	chrF	TER↓
Baseline	Gemma-3-1B-Instruct	4.99	28.87	137.41
${}^5C_1$ *LoRa*	R_1 (Key terms)	16.13	47.37	71.57
	R_2 (Syntactic)	14.77	45.31	74.13
	R_3 (Semantic)	**16.58**	**48.03**	**70.23**
	R_4 (Pragmatics)	14.45	43.36	75.36
	R_5 (Paraphrase)	15.27	44.64	74.08
${}^5C_2$ *LoRa*	R_1+R_2	15.03	45.88	74.15
	R_1+R_3	14.91	44.41	75.40
	R_1+R_4	13.30	40.25	79.89
	R_1+R_5	14.94	44.62	74.50
	R_2+R_3	15.54	46.64	72.30
	R_2+R_4	14.86	45.35	73.46
	R_2+R_5	15.89	47.07	71.39
	R_3+R_4	15.88	47.32	71.70
	R_3+R_5	**16.91**	**48.92**	**69.50**
	R_4+R_5	15.72	46.92	70.73
${}^5C_3$ *LoRa*	R_1+R_2+R_3	15.70	46.82	71.36
	R_1+R_2+R_4	14.42	44.57	74.76
	R_1+R_2+R_5	15.26	45.66	72.70
	R_1+R_3+R_4	13.72	44.97	75.57
	R_1+R_3+R_5	15.78	46.96	72.03
	R_1+R_4+R_5	14.42	44.27	75.76
	R_2+R_3+R_4	15.42	46.31	72.24
	R_2+R_3+R_5	15.52	46.52	71.62
	R_2+R_4+R_5	15.09	45.64	72.95
	R_3+R_4+R_5	**15.85**	**48.09**	**71.72**
${}^5C_4$ *LoRa*	R_1+R_2+R_3+R_4	14.45	44.75	75.57
	R_1+R_2+R_3+R_5	13.91	42.61	75.57
	R_1+R_2+R_4+R_5	13.98	44.17	76.69
	R_1+R_3+R_4+R_5	13.95	43.48	76.65
	R_2+R_3+R_4+R_5	**14.95**	**45.65**	**73.58**
${}^5C_5$ *LoRa*	Gemma all R_1 to R_5	11.64	38.35	84.13
SFT	Finetune with *LoRa*	17.55	50.00	67.41
	Full Finetune	**22.12**	**56.06**	**60.51**
Full FT R_3+R_5	Standard Inference	22.26	55.54	61.61
	Reasoning Inference	**23.86**	**57.57**	**58.55**

Table 3: Translation performance metrics for all 31 *LoRA* reasoning ablation configurations and the full-parameter fine-tuning setups. The results demonstrate that the Guided Inference approach using the $R_3 + R_5$ context yields the highest overall performance.

5.1 Phase 1: Ablation Study and the Role of Semantic Guidance

The Phase 1 ablation results show that the translation quality depends more on the type of reasoning than on the amount of reasoning added. Standard *LoRA* fine-tuning on source-target pairs achieves the best score of 17.55 BLEU, but performance does not improve monotonically as more reasoning components are introduced across ${}^5C_1$ to ${}^5C_5$. Instead, larger and more heterogeneous combinations often reduce performance, with the full five-component setting (${}^5C_5$) dropping to 11.64 BLEU and most four-component settings (${}^5C_4$) remaining below 15 BLEU points. This suggests that excessive reasoning can dilute the core translation objective and burden a compact model with unnecessary context. In contrast, compact semantic guidance proves more effective. Among single-component settings, Semantic reasoning (R_3) performs the best with 16.58 BLEU, while among pairwise settings, $R_3 + R_5$ achieves the highest score of 16.91 BLEU, showing that meaning-focused support and paraphrastic reformulation help translation more than heavier reasoning combinations.

5.2 Phase 2: Training-Inference Alignment

We next evaluate the best subset, $R_3 + R_5$, under full-parameter fine-tuning. As shown in Table 3, the standard full fine-tuning reaches 22.12 BLEU. When the model is trained with $R_3 + R_5$ but evaluated without reasoning, performance increases only slightly to 22.26 BLEU points. This small gain suggests a difference in training and testing: the model learns under reasoning-augmented inputs but cannot fully exploit that information when it is removed at test time. When the same $R_3 + R_5$ context is provided during inference, performance rises to 23.86 BLEU, with corresponding gains in chrF and TER. This result shows that reasoning is most effective when training and inference conditions remain aligned.

Model Configuration	Agri.	Tour.	Gov.	Clim.	Health	Sci&T	Judic.	Educ.	Overall
LoRA: R_3 Only	15.06	7.27	12.06	14.70	15.90	22.00	13.66	20.72	16.58
LoRA: $R_3 + R_5$	15.20	7.82	11.87	14.92	16.32	23.03	13.30	21.18	16.91
LoRA: $R_3 + R_4 + R_5$	14.89	7.50	12.29	13.99	14.84	20.43	12.58	20.16	15.85
LoRA: $R_2 + R_3 + R_4 + R_5$	12.59	7.07	10.67	14.03	14.37	19.20	11.52	19.18	14.95
LoRA: All 5	10.12	5.12	8.63	11.20	11.08	19.05	7.08	14.72	11.64
Full SFT: Standard Finetune	19.92	12.87	17.44	**21.95**	19.18	24.85	18.19	28.27	22.12
Full SFT: $R_3 + R_5$ (Normal Test)	20.20	12.26	17.73	19.90	19.68	26.31	**19.60**	28.13	22.26
Full SFT: $R_3 + R_5$ (Guided Inference)	**21.86**	**15.32**	**20.32**	21.16	**20.73**	**30.50**	19.12	**29.12**	**23.86**

Table 4: Domain-wise BLEU scores across 8 distinct evaluation domains. Guided Inference demonstrates broad robustness, yielding significant improvements in highly technical domains like Science & Technology and Governance

5.3 Domain-Wise Analysis

We further evaluate performance across eight domains, as shown in Table 4. The guided inference setup with $R_3 + R_5$ improves over the standard full fine-tuning in nearly all the domains. The largest gains appear in Science & Technology (24.85 to 30.50 BLEU), Governance (17.44 to 20.32 BLEU), and Tourism (12.87 to 15.32 BLEU). Although, Climate domain shows a small drop relative to the finetune, the overall trend indicates that compact semantic support improves robustness across diverse domains, especially in technically demanding settings.

6 Discussion

6.1 Aligning Context with Translation

The contrast across reasoning configurations reveals an important trade-off in context-augmented translation. For compact models, adding large amount of heterogeneous reasoning can weaken performance instead of improving it. We refer to this effect as *objective diffusion*: the model must allocate limited capacity to processing multiple forms of auxiliary context, which can distract from the core translation objective. In contrast, the strong performance of $R_3 + R_5$ highlights the value of *semantic anchoring*. Semantic reasoning (R_3) helps the model focus on the core meaning of the source sentence, while Paraphrase (R_5) provides a meaning-preserving reformulation that supports flexible generation. Because both the components operate at the level of semantic preservation, they align more directly with the goal of translation than broader and more heterogeneous reasoning combinations.

6.2 The Role of Guided Inference

Our results also show that training-inference alignment is critical in reasoning-augmented MT. When a model is trained with reasoning-rich inputs but evaluated without reasoning, the benefit is limited. This suggests that the model learns to rely on the additional semantic structure during training, and removing it at test time creates a difference. Reasoning guided inference addresses this issue by supplying source-derived reasoning during testing, thereby preserving the input structure seen in training. Under this setup, performance improves from 22.12 BLEU to 23.86 BLEU. This gain indicates that reasoning is most effective not as an isolated training signal, but as a consistent form of supportive context available throughout the translation process.

7 Conclusion

In this paper, we study whether structured reasoning can improve machine translation when used as supportive context for Hindi–Bengali translation. Through a complete 31-model ablation studies, we found that reasoning is not uniformly beneficial: large heterogeneous combination often degrade performance, whereas compact semantic guidance is more effective. In particular, the combination of Semantic and Paraphrase ($R_3 + R_5$) emerged as the strongest configuration. Our full-parameter experiments further showed that the benefit of reasoning depends on training-inference alignment. When reasoning is used during training but removed at test time, gains remain small. When the same $R_3 + R_5$ context is also provided during inference, performance rises from 22.12 BLEU to 23.86 BLEU points, with improvements across multiple domains. Overall, our findings suggest that reasoning can benefit compact translation models when it is semantically focused, source-grounded, and used consistently across training and inference. More broadly, the results highlight that the effectiveness of supportive context depends not on its quantity, but on how closely it aligns with the translation objective.

Acknowledgment

The authors gratefully acknowledge the COIL-D (Centre of Indian Language Data) Project under Bhashini, funded by the Ministry of Electronics and Information Technology (MeitY), Government of India, for providing the resources and computational infrastructure that supported this research.

References

Bala Das, Sudhansu, Atharv Biradar, Tapas Kumar Mishra, and Bidyut Kr. Patra. 2023. Improving multilingual neural machine translation system for indic languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 22(6), June.

Chen, Huadong, Shujian Huang, David Chiang, and Jiajun Chen. 2017. Improved neural machine translation with a syntax-aware encoder and decoder. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1936–1945, Vancouver, Canada, July. Association for Computational Linguistics.

Dabre, Raj, Bianka Buschbeck, Miriam Exel, and Hideki Tanaka. 2023. A study on the effectiveness of large language models for translation with markup. In Utiyama, Masao and Rui Wang, editors, *Proceedings of Machine Translation Summit XIX, Vol. 1: Research Track*, pages 148–159, Macau SAR, China, September. Asia-Pacific Association for Machine Translation.

Fan, Yuchun, Yongyu Mu, YiLin Wang, Lei Huang, Junhao Ruan, Bei Li, Tong Xiao, Shujian Huang, Xiaocheng Feng, and Jingbo Zhu. 2025. SLAM: Towards efficient multilingual reasoning via selective language alignment. In Rambow, Owen, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9499–9515, Abu Dhabi, UAE, January. Association for Computational Linguistics.

Gao, Hui, Jing Zhang, Peng Zhang, and Chang Yang. 2025. Consistency rating of semantic transparency: an evaluation method for metaphor competence in idiom understanding tasks. In Rambow, Owen, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio, and Steven Schockaert, editors, *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10460–10471, Abu Dhabi, UAE, January. Association for Computational Linguistics.

Gaur, Vedant and Nikunj Saunshi. 2023. Reasoning in large language models through symbolic math word problems.

Ghazaryan, Gayane, Erik Arakelyan, Isabelle Augenstein, and Pasquale Minervini. 2025. SynDARin: Synthesising datasets for automated reasoning in low-resource languages. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6459–6466, Abu Dhabi, UAE, January. Association for Computational Linguistics.

He, Jie, Tao Wang, Deyi Xiong, and Qun Liu. 2020. The box is in the pen: Evaluating commonsense reasoning in neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3662–3672, Online, November. Association for Computational Linguistics.

He, Zhiwei, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. Exploring human-like translation strategy with large language models. *Transactions of the Association for Computational Linguistics*, 12:229–246.

Jiang, Gangwei, Yahui Liu, Zhaoyi Li, Wei Bi, Fuzheng Zhang, Linqi Song, Ying Wei, and Defu Lian. 2025. What makes a good reasoning chain? uncovering structural patterns in long chain-of-thought reasoning. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6490–6514, Suzhou, China, November. Association for Computational Linguistics.

Kingma, Diederik P. and Jimmy Ba. 2017. Adam: A method for stochastic optimization.

Lee, Minjae, Youngbin Noh, and Seung Jin Lee. 2025. A testset for context-aware LLM translation in Korean-to-English discourse level translation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 1632–1646, Abu Dhabi, UAE, January. Association for Computational Linguistics.

Liu, Xuebo, Yutong Wang, Derek F. Wong, Runzhe Zhan, Liangxuan Yu, and Min Zhang. 2023. Revisiting commonsense reasoning in machine translation: Training, evaluation and challenge. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15536–15550, Toronto, Canada, July. Association for Computational Linguistics.

Nguyen, Lam and Yang Xu. 2025. Reasoning for translation: Comparative analysis of chain-of-thought and tree-of-thought prompting for LLM translation. In Zhao, Jin, Mingyang Wang, and Zhu Liu, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 259–275, Vienna, Austria, July. Association for Computational Linguistics.

Papineni, Kishore, 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.

Singh, Kshetrimayum Boynao, Ningthoujam Avichandra Singh, Loitongbam Sanayai Meetei, Ningthoujam Justwant Singh, Thoudam Doren Singh, and Sivaji Bandyopadhyay. 2023. A comparative study of transformer and transfer learning MT models for English-Manipuri. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 791–796, Goa University, Goa, India, December. NLP Association of India (NLPAI).

Singh, Kshetrimayum Boynao, Asif Ekbal, and Partha Pakray. 2025a. Evaluating IndicTrans2 and ByT5 for English–Santali machine translation using the ol chiki script. In *Proceedings of the 1st Workshop on Multimodal Models for Low-Resource Contexts and Social Impact (MMLoSo 2025)*, pages 95–100, Mumbai, India, December. Association for Computational Linguistics.

Singh, Ningthoujam Avichandra, Boynao Kshetrimayum, Ningthoujam Justwant Singh, Thoudam Doren Singh, and Shilpa Sharma. 2025b. Quantifying the impact of data scale and quality on neural machine translation for low-resource language pair. In *2025 OITS International Conference on Information Technology (OCIT)*, pages 1–6.

Singh, Kshetrimayum Boynao, Soham Bhattacharjee, Baban Gain, Asif Ekbal, and Partha Pakray. 2026a. A comparative study of fine-tuned mbart and indictrans2 for english-hindi legal machine translation. *TechRxiv*, 2026(0102).

Singh, Kshetrimayum Boynao, Soham Bhattacharjee, Baban Gain, Asif Ekbal, and Partha Pakray. 2026b. A comparative study of fine-tuned mbart and indictrans2 for english-hindi legal machine translation. *TechRxiv*, 2026(0102).

Verma, Sshubam, Mohammed Safi Ur Rahman Khan, Vishwajeet Kumar, Rudra Murthy, and Jaydeep Sen. 2025. MILU: A multi-task Indic language understanding benchmark. 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 10076–10132, Albuquerque, New Mexico, April. Association for Computational Linguistics.

Wang, Jiaan, Fandong Meng, Yunlong Liang, and Jie Zhou. 2025. DRT: Deep reasoning translation via long chain-of-thought. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 6770–6782, Vienna, Austria, July. Association for Computational Linguistics.

Yan, Shen, Leonard Dahlmann, Pavel Petrushkov, Sanjika Hewavitharana, and Shahram Khadivi. 2018. Word-based domain adaptation for neural machine translation. In *Proceedings of the 15th International Conference on Spoken Language Translation*, pages 31–38, Brussels, October 29-30. International Conference on Spoken Language Translation.

Yao, Shunyu, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Zhu, Wenhao, 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 Duh, Kevin, Helena Gomez, and Steven Bethard, editors, *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico, June. Association for Computational Linguistics.

A Appendix

B Hardware, Software, and Reproducibility

All experiments were conducted on a high-performance computing server equipped with two NVIDIA A100 GPUs, each with 80GB of memory, enabling efficient large-scale training and inference. The computational environment was configured using Python 3.13.2, along with PyTorch 2.7.1 and Transformers 4.57.3 for model implementation and experimentation. These libraries ensured compatibility with modern large language model architectures and optimized GPU acceleration.

To maintain consistency across experiments, we used fixed configurations for model loading, batching, and decoding strategies. We also experimented with a validation set during training; however, the scores remained nearly identical to training without it. Therefore, the final model was trained without a validation set to fully utilize the available training data. The inference process was optimized through batch processing and parallel execution for efficient GPU utilization. Additionally, external API-based components used for reasoning generation were managed through multiple API keys and a load-balanced request mechanism to ensure stability and scalability.

For reproducibility, we maintained detailed configurations, cached intermediate outputs such as reasoning annotations, and fixed key parameters wherever applicable. We will publicly release the complete codebase, including pre-processing scripts, inference pipelines, and configuration files, to support replication and further research. To support reproducibility, all code and datasets are publicly available on GitHub[1] and our research group webpage.[2]

[1] `https://github.com/helloboyn/RSC4MT`

[2] `https://ai-nlp-ml.github.io/resources.html`

Reasoning Generation Prompt

```
<|begin_of_text|><|start_header_id|>system<|end_header_id|>
You are an expert linguistic analyst. Analyze the given Hindi sentence and generate
5 types of reasoning.

Return ONLY one valid JSON object with exactly these keys:
"key_terms", "syntactic", "semantic", "pragmatic", "paraphrase"

Strict output rules for every value:
- Do NOT start with phrases like "The sentence", "This sentence means", or similar.
- Each value must be a single clean sentence or phrase.
- Do not add explanations outside JSON.
- Do not mention that you are analyzing.
- Avoid redundancy, filler, and obvious restatements.
- If a field is not applicable, return an empty string "".

{
  "key_terms": "Identify the most important content words (nouns, verbs, key entities,
  domain-specific terms). Focus only on meaningful lexical units essential for translation.
  Avoid full sentence repetition.",

  "syntactic": "Analyze the syntactic structure using linguistic terms: grammatical
  ↪  relations,
  subject-object-verb roles, modifiers, clause structure, and word order.
  Do not start with words like 'sentence' or 'phrase'.",

  "semantic": "Provide a precise one-sentence semantic interpretation. Capture the exact
  ↪  meaning,
  resolve ambiguity, and ensure the intended sense of words is clear.",

  "pragmatic": "Analyze contextual and pragmatic aspects in one sentence: tone, intent,
  ↪  formality,
  audience, domain, and any implied meaning or communicative purpose.",

  "paraphrase": "Generate a semantically equivalent paraphrase in one sentence, improving
  ↪  clarity
  and reducing ambiguity while preserving exact meaning."
}
<|eot_id|><|start_header_id|>user<|end_header_id|>
Sentence: "{INPUT_SENTENCE}"
Return valid JSON only.
<|eot_id|>
```

Table 5: Prompt used for generating structured linguistic reasoning from Hindi sentences. The prompt instructs the model to produce five components; key terms, syntactic structure, semantic interpretation, pragmatic context, and paraphrase used to enhance translation quality

Output Sample 1: Qualitative Comparison of Reasoning-Augmented Translation Models

Source: दोपहर के भोजन के उपरांत, शिष्टमंडल को ऐतिहासिक मेलबोर्न क्रिकेट ग्राउंड (एमसीजी) ले जाया गया।

Target: দুপুরের খাওয়ার পর, প্রতিনিধিদলকে ঐতিহাসিক মেলবোর্ন ক্রিকেট গ্রাউন্ড (এম সি জি) নিয়ে যাওয়া হয় ।

R3 LORA: দোপহর ের খাবার পর, বডিকে ঐতিহাসিক মেলবর্ন ক্রিকিং গ্রাউন্ড (এমসিজি) নিয়ে যাওয়া হয়েছিল।

R3R5 LORA: দোপহর ের খাবার পর, মাননীয়দের ঐতিহাসিক মেলবর্ন কিকিং গ্রাউন্ড (এমসিজি) নিয়ে যাওয়া হয়েছিল।

R3R4R5 LORA: দুপুর বেলার পরে, বডিমন্ডন ক্যানিংক্রিম গ্রাউন্ড (এমসিজি) - যাত্মকে নিয়ে যাওয়া হয়েছিল।

R2R3R4R5 LORA: দুপুর খাবারের পর, বডিকে ঐতিহাসিক মকম্বন ক্রিকেট গ্রাউন্ড (এমসিজি) নিয়ে যাওয়া হয়।

ALL 5 LORA: বিকেল ব্রেক করার পর, চিঠিটি ঐতিহাসিক ম chlorides ্রি ন গ্রাউন্ডে (এমসিজি) নিয়ে যাওয়া হয়।

FULL FT SFT: দুপুরের খাবারের পর, দলীয়কে ঐতিহাসিক মেলবোন ক্রিকেট মাঠ (এমসিজি) নিয়ে যাওয়া হয়েছিল।

FULL FT R3R5:

NORMAL TEST: দুপুরের খাবারের পর, প্রতিনিধিদল ঐতিহাসিক মেলবরিন ক্রিকেট গ্রাউন্ডে (এমসিজি) পৌঁছায়।

GUIDED TEST: দুপুরের খাবারের পর, প্রতিনিধিদলকে ঐতিহাসিক মেলবোর্ন ক্রিকেট গ্রাউন্ড (এমসিজি) নিয়ে যাওয়া হয়।

Table 6: The source sentence states that “After lunch, the delegation was taken to the historic Melbourne Cricket Ground (MCG)” **R3R4R5 LoRA** shows clear linguistic errors, including incorrect formations such as “বডিমন্ডন ক্যানিংক্রিম” and the non-Bengali word “যাত্মকে”. The **Full FT SFT** output contains inappropriate lexical choice (“দলীয়েক”), reducing naturalness. In Normal Test, “মেলবর্ন” is mistransliterated as “মেলবরিন”, while the Guided Test correctly produces “মেলবোর্ন”, improving accuracy. Additionally, the Guided output uses proper case marking (“প্রতিনিধিদলকে”) and construction (“নিয়ে যাওয়া হয়”), though it slightly alters voice. Overall, the **Guided Test** yields the most linguistically accurate and natural Bengali translation.

Output Sample 2: Qualitative Comparison of Reasoning-Augmented Translation Models

Source: वहीं, पंजाब स्टेट पावर कॉर्पोरेशन लिमिटेड रावी नदी पर रंजीत सागर बांध का ख्याल रखता है।

Target: অন্য দিকে, পাঞ্জাব স্টেট পাওয়ার কর্পোরেশন লিমিটেড রাভি নদীর উপর রঞ্জিত সাগর বাঁধের তত্ত্বাবধান করে।

R3 LORA: অন্যদিকে, পংজাব স্টেট পাওয়ার কর্পোরেশন লিমিটেড রাভি নদীকে রিংজাত সাগর বাঁধের দায়িত্বে রাখে।

R3R5 LORA: অন্যদিকে, পংজাব স্টেট পাওয়ার কর্পোরেশন লিমিটেড রাভি নদীকে রিংজাত সাগর বাঁধের দায়িত্বে রাখে।

R3R4R5 LORA: অন্যদিকে, পংজাব স্টেট পাওয়ার কর্পোরেশন লিমিটেড রাভি নদীকে নিয়ন্ত্রণ করে।

R2R3R4R5 LORA: অন্যদিকে, পংজাব স্টেট পাওয়ার কর্পোরেশন লিমিটেড রায়গড়□ের উপর রঞ্জিত সাগর বাঁধের যত্ন নেয়।

ALL 5 LORA: অন্যদিকে, পংজাব স্টেট পাবর কর্পোরেশন লিমিটেড রাবী নদীর উপর রিংজি সার্কিটের খেয়াল রাখে।

FULL FT SFT: একইভাবে, পাঞ্জাব স্টেট পাওয়ার কর্পোরেশন লিমিটেড রাস্তার নদীতে রঞ্জিত সাগর বাঁধের যত্ন নেয়।

FULL FT R3R5:

NORMAL TEST: একই সঙ্গে, পাঞ্জাব স্টেট পাওয়ার কর্পোরেশন লিমিটেড রাবী নদীর উপর রঞ্জিত সাগর বাঁধের যত্ন নেয়।

GUIDED TEST: একই সময়ে, পাঞ্জাব স্টেট পাওয়ার কর্পোরেশন লিমিটেড রাভি নদীর উপর রঞ্জিত সাগর বাঁধের তত্ত্বাবধান করে।

Table 7: The source sentence states that “Meanwhile, Punjab State Power Corporation Limited takes care of the Ranjit Sagar Dam on the Ravi River.“ The reference translation expresses this meaning using a formal and contextually appropriate predicate, “তত্ত্বাবধান করে”. Among the reasoning-based variants, **R3 LoRA** and **R3R5 LoRA** preserve the main entities and the overall meaning reasonably well, but their use of “দায়িত্বে রয়েছে” is slightly less precise than the reference and weakens the locative relation. In contrast, **R3R4R5 LoRA**, **R2R3R4R5 LoRA**, and **All 5 LoRA** show noticeable semantic degradation, including omission or distortion of key content words and reduced adequacy. The **Full FT SFT** and **Full FT R3R5** normal outputs recover most of the source content, but the phrase “যত্ন নেয়” is less suitable for describing institutional oversight. Overall, the **Full FT R3R5 Guided Test** output is closest to the reference “রাভি” over “রাবী”, as it best preserves the entities, relation, and formal administrative tone of the source sentence.