It Takes Two: A Dual Stage Approach for Terminology-Aware Translation

Akshat Singh Jaswal
PES University
sja.akshat@gmail.com

Abstract

This paper introduces DuTerm, a novel two-stage architecture for terminology-constrained machine translation. Our system combines a terminology-aware NMT model, adapted via fine-tuning on large-scale synthetic data, with a prompt-based LLM for post-editing. The LLM stage refines NMT output and enforces terminology adherence. We evaluate DuTerm on English-to-German, English-to-Spanish, and English-to-Russian for the WMT 2025 Terminology Shared Task. We demonstrate that flexible, context-driven terminology handling by the LLM consistently yields higher quality translations than strict constraint enforcement. Our results highlight a critical trade-off, revealing that an LLM's intrinsic knowledge often provides a stronger basis for high-quality translation than rigid, externally imposed constraints.

1 Introduction

The accurate and consistent translation of domain-specific terminology is a challenge in the field of Machine Translation and is of importance in domains such as law, medicine, and engineering, where precision is critical (Naveen and Trojovský, 2024). While modern Neural Machine Translation systems based on architectures like the Transformer have achieved remarkable fluency and quality on general text, their performance in terminology-constrained texts remains a critical area for improvement (Vaswani et al., 2023; Bahdanau et al., 2016; Johnson et al., 2017). This issue is particularly relevant given findings of recent WMT shared tasks, which have consistently highlighted the need for systems that can effectively handle domain-specific vocabulary (Post, 2018). The WMT 2025 Terminology Shared Task (Semenov et al., 2025) provides a focused platform to evaluate MT systems ability to handle domain-specific terminology under controlled conditions across multiple language pairs: English to German, English to Spanish, and English to Russian.

Previous research into terminology-constrained MT can be broadly categorized into two main approaches: inference-time methods and training-time methods. Inference-time approaches incorporate terminology constraints directly into the decoding process, often through techniques like constrained beam search or by re-ranking n-best lists of candidate translations (Zhang et al., 2023). While these methods are highly effective at enforcing constraints, they can be computationally expensive and may compromise the overall fluency and grammatical correctness of the output by forcing the model to generate awkward or unnatural phrases. Recent work has explored ways to make these methods more efficient, but the trade-off between enforcing terminology constraints and fluency remains a key consideration (Moslem et al., 2023).

Alternatively, training-time methods aim to teach models how to handle terminology constraints by integrating the terminology information into the training data itself. This is commonly done through the use of special tags that surround the terms to be translated (Dinu et al., 2019). This approach allows the model to learn how to produce more natural and grammatically correct output, but it provides no guarantee that all constraints will be respected during inference (Susanto et al., 2020).

We present DuTerm, a two-stage architecture that addresses these limitations by combining the strengths of both training-time and inference-time methodologies. We recognize that terminology-constrained translation is not merely a lexical substitution problem but requires a deeper understanding of linguistic context, especially when dealing with the complex morphology of languages like German and Russian. Our system is specifically designed to tackle the multifaceted evaluation framework of the WMT 2025 shared task.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 1102–1108
November 8–9, 2025.

2 Method

2.1 Terminology-Aware Neural Machine Translation

Overview We develop a terminology aware MT model via large-scale, tagged synthetic data and targeted fine-tuning. The pipeline: extract and analyze terminology, generate tagged, context-rich parallel data (single-term and multi-term) and standardize tags and ensure annotation consistency. We also quality-filter with COMET_{QE} (Rei et al., 2022) and deduplicate , this finally adapts a multilingual NMT model with parameter efficient fine-tuning .

Terminology Extraction and Analysis We parse the WMT 2025 dev files for English→German/Spanish/Russian to build bilingual terminology dictionaries. The dictionaries typically exceed 1,000 unique pairs per direction. We track terms and occurrences using `repetition_ids`. We also use the LLM to generate more terms similar to the terms provided in the dictionaries.

Synthetic Data Generation We use GPT-4o (OpenAI, 2024) to create parallel sentences that naturally embed required terms and explicitly insert boundary tags (`[TERM]...[/TERM]`) on both source and target. There are two modes we use to generate these parallel sentences

Single-term mode: generates sentence pairs containing exactly one term instance per sentence.

Multi-term mode: randomly selects 2–3 term pairs to appear together, teaching co-occurrence handling and disambiguation.

We employ temperature sampling (0.3–0.7), concurrent generation, and strict parsing to yield well-formed bilingual pairs.

Tag Standardization and Quality Filtering A re-tagging pass enforces consistent annotation, longest-first matching prevents partial shadowing, case-insensitive detection with original case preservation, and inverse mapping ensures symmetric target-side tagging. Each pair is scored by COMET_{QE}. We deduplicate on the source side and keep only high-confidence items using a conservative threshold (0.85-0.9) depending on the language, typically retaining 60–70% of outputs, yielding ∼10k–15k pairs per language direction.

Multilingual Model Adaptation For the foundation translation model, we select NLLB-200 3.3B, a state-of-the-art multilingual neural machine translation model with demonstrated strong performance across our target languages (Team et al., 2022). This model provides robust baseline capabilities while supporting the specialized terminology handling adaptations we require.

We extend the model's vocabulary with terminology markup tokens to ensure atomic treatment of terminology annotations. This prevents subword tokenization from fragmenting our special markup, ensuring that terminology boundaries are consistently preserved during training and inference.

The training process employs several optimization strategies designed for stable, effective adaptation. The process also combines filtered datasets from all three target languages, creating unified multilingual adaptation that benefits from cross-lingual transfer.

2.2 LLM-Based Post-Editing

Overview An LLM refines the NMT output given the source sentence and required term pairs, enforcing strict terminology adherence while improving fluency and morphology.(Raunak et al., 2023)

Post-Editing Procedure We use prompts that present the source, translation, and provide explicit source to target term mappings. The LLM is instructed to preserve meaning, apply the exact target terms, maintain tags where required, and improve readability without paraphrasing away constraints. The LLM we choose to use is GPT-4o (OpenAI, 2024) due to its combination of high translation quality and relatively lower price.

Terminology-Aware Processing *Dynamic resolution*: per-input selection of proper/random/no-term constraints from reference terminology databases with whitespace-normalized matching. *Mode-adaptive behavior*: when constraints exist, the LLM must enforce them; otherwise it performs quality-only edits while being sensitive to technical terms. *Constraint satisfaction*: explicit mappings and formatting rules are included in the prompt; outputs must preserve required terminology and markup.

Quality Assurance and Robustness We run the LLM at low temperatures (0.3) for deterministic edits. Each hypothesis is validated for format, tag integrity, and constraint satisfaction before acceptance with a pre-existing parser. We verify filename

schemas, presence of all terminology modes per language pair, and JSONL structure. We assess quality with COMET_{QE} (after tag stripping) and compute terminology preservation via exact-match checks on required terms. This ensures reliability of final outputs.

3 Results

We evaluate the system using three complementary metrics used by the WMT organizers: BLEU for overall translation adequacy, chrF2++ for character-level fluency and robustness, and terminology success rates (proper and random) to directly measure constraint satisfaction (Papineni et al., 2002; Popović, 2015). Results are reported for English→German (DE), English→Spanish (ES), and English→Russian (RU) across three terminology strategies: *noterm*, *proper*, and *random*.

Table 1 summarizes the findings. Several clear patterns emerge:

1. **Strict terminology enforcement (*proper*)** achieves the highest BLEU and chrF2++ across all languages (48.06 for DE, 58.51 for ES, 35.80 for RU), indicating improved lexical precision and sentence-level quality when constraints are respected. It also yields near-perfect proper terminology success rates ($\geq$0.97).

2. **Unconstrained translation (*noterm*)** consistently underperforms, producing the lowest BLEU and chrF2++ values across languages (e.g., 38.24 BLEU in DE and 27.88 in RU). While fluency remains reasonable, failure to enforce constraints leads to poor terminology precision.

3. **Random terminology enforcement** produces intermediate BLEU/chrF2++ but near-perfect random-term success rates ($\sim$0.98). This highlights that while the model can force arbitrary terminology, doing so compromises contextual appropriateness.

4. **Language-specific trends** align with expectations: Spanish shows the highest overall scores, reflecting its structural similarity to English. Russian shows the widest gap between *proper* and *noterm*, emphasizing the difficulty of morphology-rich languages for terminology control.

Overall, these results demonstrate that while strict enforcement maximizes terminology accuracy and boosts surface-level quality metrics, it can occasionally reduce flexibility. In contrast, unconstrained approaches produce more natural translations but risk terminology inconsistency.

4 Conclusion

This paper presents our approach to the terminology shared task, focusing on English to German, Spanish, and Russian translation directions. Our system leverages LLMs to improve existing translations with varying terminology handling strategies. Our results demonstrate that allowing the LLM to flexibly handle terminology often yields higher translation quality than strict terminology enforcement. These findings highlight the potential of prompt-based LLM systems for technical and business translation tasks, and provide insights into effective strategies for terminology management in neural translation workflows. The intution behind why this approach works so well is that NMTs often excel at strict word-level translations however they can struggle with context-dependent nuances. Our approach leverages post-editing with a LLM on top of the initial NMT outputs. By starting from a reliable NMT translation, the post-editing model receives a structured, partially correct target sentence, which allows it to focus on higher-level improvements resolving ambiguities, adjusting word order, and refining context. This guided refinement is often more effective than translating directly from scratch with an LLM or NMT which must simultaneously handle term accuracy and contextual fluency.

For future work, exploring adaptive learning mechanisms that integrate terminology dynamically, rather than relying on static prompts, could enhance robustness across domains and languages. End-to-end or memory-augmented architectures that maintain consistency across sentences and documents hold promise for more coherent outputs. Expanding evaluations to other language models and diverse, domain-specific corpora would help validate the approach's generalizability and reveal domain-dependent challenges. Incorporating hybrid strategies, such as combining prompt guidance with fine-tuning or reinforcement learning, and enabling user-driven interaction for terminology control could further improve usability and accuracy. Together, these directions offer a pathway toward more flexible, context-aware, and widely applicable terminology-aware translation systems.

Limitations

Our approach, based on a prompt-driven framework, faces several limitations. It depends heavily on carefully crafted prompts, which may not gen-

Lang	Type	BLEU	chrF2++	Prop. SR	Rand. SR
DE	noterm	38.24	62.61	0.43	0.69
	proper	48.06	70.74	0.98	0.73
	random	43.77	67.22	0.48	0.99
ES	noterm	45.98	67.05	0.47	0.73
	proper	58.51	76.08	0.99	0.78
	random	53.28	72.05	0.49	0.98
RU	noterm	27.88	55.29	0.39	0.69
	proper	35.80	63.57	0.98	0.72
	random	32.25	59.85	0.42	0.99

Table 1: Evaluation results for English→German (DE), English→Spanish (ES), and English→Russian (RU) across three terminology handling strategies. Metrics include BLEU, chrF2++, and terminology success rates (proper and random).

eralize well across domains, languages, or model architectures. The sequential processing of terminology matching and translation refinement limits the system's ability to adaptively enforce terminology constraints. Furthermore, operating at the sentence level overlooks opportunities for document-level consistency and context-aware terminology usage, which are crucial in practical translation tasks. Our evaluation, conducted solely on GPT-4o (OpenAI, 2024), restricts the generalizability of findings, and focusing on technical and business domains may not capture challenges present in specialized fields like medical or legal translation. Additionally, while COMET_{QE} ,BLEU ,chrF++ provide automated scalability, it may not fully reflect terminological precision and contextual appropriateness, suggesting the need for complementary evaluation methods that include human judgment.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2016. Neural machine translation by jointly learning to align and translate.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3893–3898, Florence, Italy. Association for Computational Linguistics.

Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation.

Yasmin Moslem, Gianfranco Romani, Mahdi Molaei, John D. Kelleher, Rejwanul Haque, and Andy Way. 2023. Domain terminology integration into machine translation: Leveraging large language models. In *Proceedings of the Eighth Conference on Machine Translation*, pages 902–911, Singapore. Association for Computational Linguistics.

Palanichamy Naveen and Pavel Trojovský. 2024. Overview and challenges of machine translation for contextually appropriate translations. *iScience*, 27(10):110878.

OpenAI. 2024. Gpt-4o system card.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Vikas Raunak, Amr Sharaf, Yiren Wang, Hany Hassan Awadallah, and Arul Menezes. 2023. Leveraging gpt-4 for automatic translation post-editing.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 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, Abu Dhabi,

United Arab Emirates (Hybrid). Association for Computational Linguistics.

Kirill Semenov, Xu Huang, Vilém Zouhar, Nathaniel Berger, Dawei Zhu, Arturo Oncevay, and Pinzhen Chen. 2025. Findings of the WMT25 Terminology Translation Task: Terminology is Useful Especially for Good MTs. In *Proceedings of the Tenth Conference on Machine Translation*, Suzhou, China. Association for Computational Linguistics.

Raymond Hendy Susanto, Shamil Chollampatt, and Liling Tan. 2020. Lexically constrained neural machine translation with levenshtein transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3536–3543, Online. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

Huaao Zhang, Qiang Wang, Bo Qin, Zelin Shi, Haibo Wang, and Ming Chen. 2023. Understanding and improving the robustness of terminology constraints in neural machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6029–6042, Toronto, Canada. Association for Computational Linguistics.

A Prompts

We include below the full prompts used in our experiments for reproducibility.

A.1 Single-Term Prompt

```
Generate {n} professional, domain-specific English-({target_lang}) bilingual sentence pairs for
    terminology translation.
The term pair to use is: {source_term}\(EN) : \"{target_term}\ ({target_lang})
Requirements:
- Each sentence pair must be natural, fluent, and contextually appropriate for IT or financial
    domains.
- Include the term exactly once per sentence.
- Wrap the term with [TERM] and [/TERM] in both the English and ({target_lang}) sentences.
- Ensure accurate translation and alignment of the term.
Format:
EN: [sentence]
{target_lang}: [sentence]
Output exactly {n} such pairs.
```

Listing 1: Prompt template for generating bilingual sentence pairs with a single terminology constraint.

A.2 Multi-Term Prompt

```
Generate {n} professional, domain-specific English-({target_lang}) bilingual sentence pairs for
    terminology translation.
Use ALL of the following term pairs in each sentence pair:\n{terms_str}
Requirements:
- Each sentence pair must be natural, fluent, and contextually appropriate for IT or financial
    domains.\n"
- Include each term exactly once per sentence.
- Wrap each term with [TERM] and [/TERM] in both the English and ({target_lang}) sentences.\n"
- Ensure accurate translation and alignment of the terms.
Format:
EN: [sentence]
{target_lang}: [sentence]
Output exactly {n} such pairs.
```

Listing 2: Prompt template for generating bilingual sentence pairs with multiple terminology constraints.

A.3 Post-Editing with Terminology

```
As an expert English-{target_lang} translator specializing in technical and business documentation,
    improve this {target_lang} translation.

SOURCE (English): {source}

CURRENT TRANSLATION ({target_lang}): {translation}

REQUIRED TERMINOLOGY (English: {target_lang}): {term_str}

YOUR TASK:
1. Ensure all required terminology is correctly used
2. Maintain the same meaning as the source text
3. Ensure natural, fluent {target_lang} that sounds like native content
4. Preserve formatting, numbers, and special characters
5. Match the tone and register of the original text

Return ONLY the improved {target_lang} translation with no explanations, notes, or other text.
```

Listing 3: Prompt for post-editing with explicit terminology mappings.

A.4 Post-Editing without Terminology

```
As an expert English-{target_lang} translator specializing in technical and business documentation,
    improve this {target_lang} translation.

SOURCE (English): {source}

CURRENT TRANSLATION ({target_lang}): {translation}

Note: There may be important terminology in the source text that should be translated precisely and
    consistently. Please ensure any technical or business terms are rendered correctly in {
    target_lang}.

YOUR TASK:
1. Enhance the translation for fluency and accuracy
2. Maintain the same meaning as the source text
3. Ensure natural, fluent {target_lang} that sounds like native content
4. Preserve formatting, numbers, and special characters
5. Match the tone and register of the original text

Return ONLY the improved {target_lang} translation with no explanations, notes, or other text.
```

Listing 4: Prompt for post-editing without explicit terminology guidance.