HW-TSC's Submission to the WMT26 Terminology Translation Shared Task

Boqi Huang, Daimeng Wei, Jiaxin Guo, Yuanchang Luo, Hengchao Shang, Zongyao Li, Jinlong Yang, Zhanglin Wu, Yu He, Xiaoqing Lan
Huawei Translation Service Center, Beijing, China
{huangboqi1,weidaimeng,guojiaxin1,luoyuanchang1,shanghengchao, lizongyao,yangjinlong7,wuzhanglin2,heyu97,lanxiaoqing}@huawei.com

Abstract

We describe HW-TSC's submissions to both tracks and all language pairs of the WMT26 Terminology Translation Shared Task. Our primary system uses Qwen3.7-Max for source-term extraction, bilingual term alignment, and document translation. In Track 1, it matches the current document or translation chunk against the supplied dictionary and places at most 200 applicable term pairs in the translation prompt. In Track 2, it first induces a bilingual glossary from the seed bitexts and then reuses the Track 1 translation module. Induction combines open source-term extraction with seed-source expression matching, followed by LLM alignment to the seed targets. We optimize the extraction prompt with SkillOpt (Yang et al., 2026) on a train/validation/test split of the WMT25 English–Chinese finance data and use Qwen3.7-Max to rewrite it for directions without labeled optimization data. Extraction recall rises from 0.6250 to 0.6344 on validation and from 0.6451 to 0.6674 on the held-out test split. On the same held-out test documents with Qwen3.7-Max, glossary induction raises TSR from 0.5212 (no glossary) to 0.7910 with the initial prompt and to 0.8026 after SkillOpt, with chrF++ of 57.45, 67.05, and 67.36; the official glossary yields TSR 0.9019 and chrF++ 63.94.

1 Introduction

Specialized translation requires more than selecting the right target lemma. A system must reuse terminology consistently across a document while realizing prescribed terms grammatically in context, a particular concern for lower-resource, morphologically rich target languages. Earlier WMT terminology tasks established evaluations that combine general translation quality with term-centered criteria (Alam et al., 2021; Semenov et al., 2023, 2025). Methods have included terminology-aware training, source annotations, constrained decoding, and, more recently, glossary-conditioned LLM generation and quality–terminology reranking (Dinu et al., 2019; Ailem et al., 2021; Bergmanis and Pinnis, 2021; Guttmann et al., 2025; Zhu et al., 2025).

WMT26 changes the operational setting in two important ways (Organizers, 2026). Both tracks are document-level. Track 1 supplies corpus-level dictionaries and requires no-, proper-, and random-terminology modes. Track 2 supplies seed bitexts rather than a dictionary, so the system must recover or otherwise exploit their terminology before translating related documents. The directions are Spanish→Basque and English→Polish in both tracks, plus Traditional Chinese→English finance in Track 2. Evaluation jointly considers overall quality, terminology success rate (TSR), document-level consistency, and grammaticality of term realizations; the organizers frame comparison in terms of the quality–terminology Pareto frontier.

Our submission is an LLM pipeline for this setting. It contributes (i) one translation module that locally matches an optional glossary and admits at most 200 pairs per document or chunk, and (ii) a Track 2 induction stage that combines two source-term discovery channels with seed-grounded target alignment. For Channel A extraction, we use SkillOpt (Yang et al., 2026), a text-space skill optimizer that treats a natural-language skill/prompt document as the trainable external state of a frozen language model, proposes bounded edits from scored rollouts, and accepts an edit only when it improves a held-out validation score. We apply SkillOpt to the WMT25 English–Chinese finance setting and transfer the resulting prompt to the remaining WMT26 directions with Qwen3.7-Max.

2 Task Coverage

HW-TSC submitted to both tracks and all available directions (Table 1). Track 1 provides source documents and domain-level proper and random dictionaries; the same source documents are also translated without terminology. Track 2 provides

Proceedings of the Eleventh Conference on Machine Translation, pages 1801–1806
October 28-29, 2026 ©2026 Association for Computational Linguistics

Track	Direction	Domain	Term resource
1	Es→Eu	Eng./Tech.	Supplied
1	En→Pl	Eng./Tech.; Med.	Supplied
2	Es→Eu	Eng./Tech.	Induced
2	En→Pl	Eng./Tech.; Med.	Induced
2	Zh-Hant→En	Finance	Induced

Table 1: Submission coverage. Track 1 uses the supplied proper or random dictionary, or omits it in the no-term mode. Track 2 derives a glossary from seed bitexts.

terminology-rich parallel seeds and related source documents, with no explicit dictionary (Organizers, 2026).

3 System Description

Figure 1 separates Track 2 glossary induction from the translator shared by both tracks. Qwen3.7-Max (Yang et al., 2025) is the primary backbone for extraction, alignment, prompt transfer, and translation. Qwen3.7-Max is also used to rewrite the optimized extraction prompt for new language/domain settings.

3.1 Shared Translation Module

Local glossary matching. The supplied Track 1 dictionaries and Track 2 induced glossaries are corpus-level resources and can contain entries irrelevant to a particular input. Before translation, we match source entries against the current source span and retain only applicable pairs. We cap the matched glossary at 200 pairs per document or chunk to control prompt length and reduce interference from irrelevant terminology. If more than 200 entries match, the implementation truncates the matched set to 200 using its selection heuristic. A short document receives at most 200 matched pairs, while a long document is first divided into chunks and each chunk is matched independently to obtain its own set of at most 200 pairs. Because matching is local to the current span, the effective terminology budget scales with the number of chunks rather than requiring a single unbounded glossary for an entire long document.

Prompted translation. The prompt specifies the source and target languages, domain, source span, and matched `source => target` pairs. It asks the model to use applicable target renderings consistently while adapting morphology and agreement to the target context. Proper and random Track 1 modes use the same matching procedure and prompt template; only dictionary content changes. Local matching is applied uniformly to both proper and random terminology settings to filter out entries that do not occur in the current source span. The no-term mode removes both the glossary block and glossary-specific instruction. Long-document chunking is invoked only when required by the input length.

The translation prompt template is as follows:

```
You are a professional translator.
Translate the following SOURCE document
from {SRC} to {TGT} in the {DOMAIN}
domain.

Requirements:
- Use the GLOSSARY entries consistently
throughout the translation.
- Adapt morphology and agreement as
required by {TGT} grammar.
- Preserve document structure,
formatting, numbers, and codes.
- Output only the translation.

GLOSSARY (source => target):
{term1 => translation1
term2 => translation2
...
}

SOURCE:
{document_or_chunk}
```

3.2 Track 2 Glossary Induction

Track 2 turns the supplied seed bitexts into a bilingual glossary before calling the shared translator. The pipeline first discovers source candidates through two channels, merges and filters them, and then asks Qwen3.7-Max to align retained candidates to target expressions evidenced in the seed translations.

3.2.1 Channel A: open source-term extraction

Channel A asks Qwen3.7-Max to extract source-language domain terms from the document to translate, without proposing translations. We optimize this extraction instruction offline with SkillOpt (Yang et al., 2026). As introduced above, SkillOpt keeps the target model frozen and updates only the skill document under validation gating. We optimize the Channel A extraction prompt on the WMT25 English–Chinese finance split using source-term recall as the validation objective, while keeping the remaining pipeline unchanged. The resulting skill replaces the hand-written extraction prompt used at inference for that setting. The optimization data are the WMT25 Track 2 English–Chinese finance data (Semenov et al., 2025), partitioned into train, validation, and held-out test subsets. Training and validation information is used

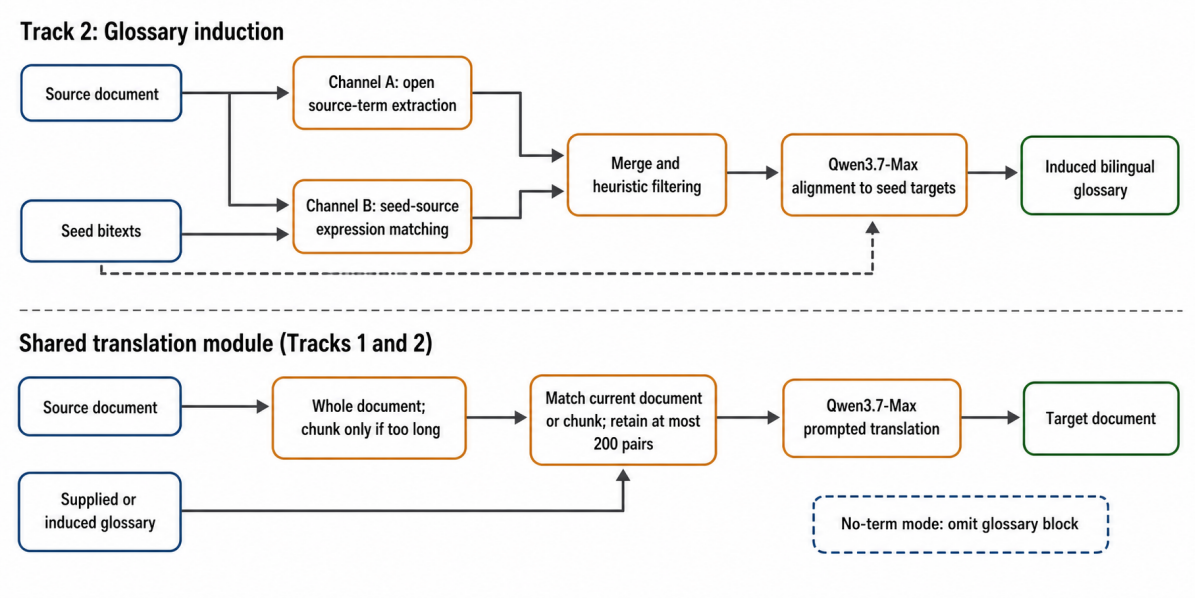

Figure 1: HW-TSC pipeline. Track 2 first induces a seed-grounded bilingual glossary. The shared translator matches only entries relevant to the current document or chunk and places at most 200 pairs in the prompt.

during prompt optimization and selection; the held-out test subset is reserved for the reported final comparison.

The optimized prompt is used for the English/Chinese finance setting. For the remaining WMT26 directions, Qwen3.7-Max adapts the optimized prompt using the target language pair, domain, and seed examples, without additional labeled optimization.

The initial prompt broadly prioritizes finance-related terms, proper names, complete multiword expressions, translation-sensitive core terms, and both abbreviations and their long forms. SkillOpt retains this task definition but makes the search procedure and span policy substantially more explicit (Table 2). In particular, it instructs the model to exhaust four candidate classes in sequence and to retain independently useful nested terms in addition to complete expressions. These edits mainly change which source spans enter the candidate pool: denser coverage of abbreviations, short core terms, and nested constituents tends to yield a slightly larger, more complete pseudo-glossary after alignment. Section 4 measures that effect as extraction recall and as downstream TSR/chrF++ under an otherwise fixed translator.

Aspect	Initial prompt	SkillOpt prompt
Search order	Broad priorities listed in parallel	Four mandatory passes: abbreviations; 1–2 word core terms; complete expressions/entities; other terms
Coverage	Important core terms and long/short forms	Exhaustive emphasis on abbreviations and short domain-core terms
Span policy	Prefer complete multiword entries; avoid redundant variants	Also emit independent nested terms and coordinated components
Surface form	No dedicated rule	Preserve quoted, punctuated, versioned, or year-qualified names; retain useful core forms separately
Capacity policy	General relevance criteria	If limited, favor abbreviations and short core terms over descriptive phrases; never fill with generic words

Table 2: Qualitative changes introduced by SkillOpt between the initial and optimized prompts.

3.2.2 Channel B: seed-source expression matching

Channel B compares the source document with the source side of the seed bitexts. It extracts expressions that are identical or differ only slightly in form, thereby recovering candidates directly connected to bilingual evidence. This channel complements open extraction: Channel A can propose domain terms from the document, whereas Channel B emphasizes expressions attested in the seeds. The exact and near-match decisions are made by Qwen3.7-Max, which identifies exact and near-

Stage	Model	Output
Prompt transfer	Qwen3.7-Max	Rewritten extraction prompt
Open extraction	Qwen3.7-Max	Source-term candidates
Seed matching	Qwen3.7-Max	Shared/near expressions
Alignment	Qwen3.7-Max	Bilingual term pairs
Translation	Qwen3.7-Max	Target span

Table 3: Model roles in the primary submission pipeline.

matching expressions against the seed sources.

3.2.3 Merge, filtering, and bilingual alignment

We merge both candidate lists and apply heuristic filtering to remove duplicates and candidates that are not usable as terms. Qwen3.7-Max then inspects the relevant seed bitexts and identifies the seed-attested target realization for each retained source candidate. Candidates for which a target cannot be grounded in the seed translations are discarded. The surviving source–target pairs form the pseudo-glossary supplied to the shared module in Section 3.1, where they are matched again against the current document or chunk and capped at 200.

Evidence boundaries. Both discovery channels generate source-side candidates. Channel A receives the document to translate and emits source candidates; Channel B compares that document with seed sources and emits shared or lightly varying source expressions. Target-side seed translations are introduced only during bilingual alignment, where each candidate is mapped to a seed-attested target expression; candidates without sufficient bilingual evidence are discarded.

4 Implementation and Evaluation

4.1 Implementation Summary

The primary submission uses Qwen3.7-Max for prompt transfer, Channel A extraction, Channel B matching and bilingual alignment, and final translation. Term discovery operates over the document, while chunking is a translation-time length safeguard. For every translation call, matching is rerun on the actual source span and limits the terminology block to at most 200 pairs.

Table 3 summarizes model responsibilities.

4.2 SkillOpt Extraction Recall

Table 4 reports source-term extraction recall before and after SkillOpt on our WMT25 English–Chinese finance split. Recall increases by 0.0094 on the validation subset and by 0.0223 on the held-out test subset. The additional candidates that survive filtering and seed-grounded alignment become extra glossary constraints for the shared translator in Section 3.1.

Split	Initial prompt	SkillOpt prompt
Validation	0.6250	0.6344
Held-out test	0.6451	0.6674

Table 4: Source-term extraction recall on our split of the WMT25 Track 2 English–Chinese finance data. Validation participates in prompt selection; the test subset is held out.

4.3 Downstream TSR and chrF++

We further evaluate Qwen3.7-Max translation on the same held-out WMT25 English–Chinese finance test documents, reporting terminology success rate (TSR) and chrF++. Table 5 compares four terminology conditions under the shared translation module: no glossary; a glossary induced by the initial Channel A prompt followed by seed-grounded alignment; a glossary induced after SkillOpt optimization of Channel A; and the official shared-task glossary.

Relative to the no-glossary baseline, the initial induced glossary raises TSR from 0.5212 to 0.7910 and chrF++ from 57.45 to 67.05. SkillOpt yields a further gain to TSR 0.8026 and chrF++ 67.36, consistent with the higher extraction recall in Table 4: a modestly richer seed-grounded glossary improves both term realization and surface overlap with the reference. The official glossary attains the highest TSR (0.9019) but a lower chrF++ (63.94) than either induced glossary. One possible explanation is that denser exact-term constraints can displace otherwise fluent phrasings or encourage locally awkward insertions, which would lower character n-gram overlap even when term hits improve. By contrast, induced glossaries are sparser and filtered by seed evidence; they recover much of the terminology benefit while leaving the translator more freedom on non-term spans, yielding higher chrF++ at a lower TSR. These results highlight a trade-off between strict terminology realization and overall surface similarity. These results are development measurements on our WMT25 English–Chinese split rather than official WMT26 scores across all submitted directions.

Terminology condition	TSR	chrF++
No glossary	0.5212	57.45
Induced (initial prompt)	0.7910	67.05
Induced (SkillOpt prompt)	0.8026	67.36
Official glossary	0.9019	63.94

Table 5: Qwen3.7-Max TSR and chrF++ on our held-out WMT25 Track 2 English–Chinese finance test split under different terminology conditions.

5 Design Considerations

The unified interface reduces Track 2 to obtaining a glossary that the Track 1 translator can consume. This simplifies implementation and makes the cap and matching behavior identical across tracks, but it also creates an error path from source-term discovery through alignment to translation. Seed-grounded alignment is intended to limit unsupported target constraints; Table 5 shows that induced glossaries already improve TSR and chrF++ over a no-glossary baseline on our WMT25 development split.

The 200-pair cap keeps each translation call within a manageable prompt budget while preserving local relevance. With per-chunk rematching, long documents are not forced to compress an entire corpus dictionary into one call; each span only competes for its own 200-pair allotment. For Basque and Polish, a further consideration arises between preserving the prescribed target lemma and producing a grammatically licensed surface form. The prompt therefore requests morphological adaptation rather than literal string copying, consistent with prior work on lemma-level terminology control (Bergmanis and Pinnis, 2021) and with the shared task's grammaticality criterion (Organizers, 2026).

Retrieval and truncation. Local matching reduces exposure to irrelevant dictionary entries. When the 200-pair cap is active, the matched set is truncated by an implementation heuristic. In the random terminology setting, local matching prevents unmatched dictionary entries from entering the translation prompt, thereby reducing potential interference.

Induction. Channel A can propose domain expressions that are not recoverable from the seeds, while Channel B can capture surface variations. The seed-grounded aligner requires evidence in the seed translations and can discard unsupported candidates. On our WMT25 English–Chinese finance split, SkillOpt acts upstream of this aligner by changing the source-candidate distribution; the downstream TSR/chrF++ gains in Table 5 are therefore an indirect effect of better Channel A coverage under a fixed translator.

Document segmentation. Matching every chunk against the same glossary makes prescribed pairs available wherever their source form is detected. Chunks are translated in separate model calls, so discourse choices and term realizations may vary across boundaries.

6 Conclusion

HW-TSC participated in both WMT26 terminology tracks and all language pairs with a Qwen3.7-Max-centered document translation system. Track 1 locally matches the supplied dictionary; Track 2 combines open extraction and seed-source matching, aligns candidates to seed targets, and feeds the induced glossary to the same translator. Every translation call receives at most 200 locally matched term pairs, with chunking only for overlong documents. On our WMT25 English–Chinese finance split, SkillOpt improves held-out source-term extraction recall from 0.6451 to 0.6674 and, after glossary induction and translation, raises TSR/chrF++ from 0.7910/67.05 (initial prompt) to 0.8026/67.36, compared with 0.5212/57.45 without a glossary.

Limitations

Quantitative analysis in this paper focuses on a partition of the WMT25 English–Chinese finance data; we do not report corresponding TSR/chrF++ breakdowns for the Basque, Polish, or Traditional Chinese WMT26 directions, nor official shared-task leaderboard scores. Validation data participates in SkillOpt prompt selection. Prompt rewrites used to transfer the SkillOpt skill across languages are not evaluated on labeled extraction sets. We likewise do not provide a measured comparison of Track 1 proper versus random dictionaries; the discussion of random-mode noise is qualitative and relies on local matching as a filter. API model versions and decoding settings may affect reproducibility. The full production prompts are not reproduced; Table 2 records their principal methodological differences.

Acknowledgments

We thank the WMT26 terminology organizers for the task design and data, and colleagues at HW-TSC for discussion and engineering support.

References

Melissa Ailem, Jingshu Liu, and Raheel Qader. 2021. Encouraging neural machine translation to satisfy terminology constraints. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1450–1455, Online. Association for Computational Linguistics.

Md Mahfuz Ibn Alam, Ivana Kvapilíková, Antonios Anastasopoulos, Laurent Besacier, Georgiana Dinu, Marcello Federico, Matthias Gallé, Kweonwoo Jung, Philipp Koehn, and Vassilina Nikoulina. 2021. Findings of the WMT shared task on machine translation using terminologies. In *Proceedings of the Sixth Conference on Machine Translation*, pages 652–663, Online. Association for Computational Linguistics.

Toms Bergmanis and Mārcis Pinnis. 2021. Facilitating terminology translation with target lemma annotations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3105–3111, Online. Association for Computational Linguistics.

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 3063–3068, Florence, Italy. Association for Computational Linguistics.

Kamil Guttmann, Adrian Charkiewicz, Zofia Rostek, Mikołaj Pokrywka, and Artur Nowakowski. 2025. Laniqo at WMT25 terminology translation task: A multi-objective reranking strategy for terminology-aware translation via Pareto-optimal decoding. In *Proceedings of the Tenth Conference on Machine Translation*, pages 1276–1283, Suzhou, China. Association for Computational Linguistics.

WMT Terminology Organizers. 2026. WMT 2026 terminology translation shared task. `https://www2.statmt.org/wmt26/terminology.html`. Accessed 2026-08-07.

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*, pages 554–576, Suzhou, China. Association for Computational Linguistics.

Kirill Semenov, Vilém Zouhar, Tom Kocmi, Dongdong Zhang, Wangchunshu Zhou, and Yuchen Eleanor Jiang. 2023. Findings of the WMT 2023 shared task on machine translation with terminologies. In *Proceedings of the Eighth Conference on Machine Translation*, pages 663–671, Singapore. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Yifan Yang, Ziyang Gong, Weiquan Huang, Qihao Yang, Ziwei Zhou, Zisu Huang, Yan Li, Xuemei Gao, Qi Dai, Bei Liu, Kai Qiu, Yuqing Yang, Dongdong Chen, Xue Yang, and Chong Luo. 2026. SkillOpt: Executive strategy for self-evolving agent skills. *Preprint*, arXiv:2605.23904.

Lichao Zhu, Maria Zimina-Poirot, Cristian Valdez, and Stephane Patin. 2025. Fine-tuning NMT models and LLMs for specialised EN-ES translation using aligned corpora, glossaries, and synthetic data: MULTITAN at WMT25 terminology shared task. In *Proceedings of the Tenth Conference on Machine Translation*, pages 1284–1291, Suzhou, China. Association for Computational Linguistics.