

SCIR-TG-MT’s Submission for the WMT 2026 General Machine Translation Task

Xiayu Cao, Baohang Li, Zekun Yuan, Zekai Ye, Xiaocheng Feng

Harbin Institute of Technology

{xycao, baohangli, zkyuan, zkye, xcfeng}@ir.hit.edu.cn

Abstract

This paper presents SCIR-TG-MT, our primary submission (ID 44) to the unconstrained track of the WMT 2026 General Machine Translation shared task. The system covers ten translation directions using DeepSeek-V4-Pro and STRIDE (Structured Translation with Risk-Informed Directional Execution), an OpenCode translation skill. STRIDE converts source-side evidence and task instructions into a translation contract, an execution route, generation constraints, and selected checks. On five WMT25 development directions, STRIDE raises average SacreBLEU from 31.24 to 31.80 and COMET-22 from 83.61 to 84.17. A cumulative ablation on three directions separates the effects of routing, controls, and validation.

1 Introduction

The WMT 2026 General MT task covers multiple languages, domains, instructions, and document formats. In addition to plain text, the data contain terminology constraints, style requirements, and structured formats including JSON, HTML, and CSV (WMT 2026 General Machine Translation Task Organizers, 2026). A single direct-translation prompt handles these records uniformly even though their binding requirements differ.

In our WMT25 development runs, direct translation showed recurring problems with structural fidelity, terminology consistency, pragmatic intent, and document completeness. STRIDE uses the source form, task instruction, and direction profile to build a translation contract and validation plan. Plain local inputs follow a short direct path and normally proceed to finalization without validation. A short-path output is validated only when independent record-level evidence selects a validator. Structure-sensitive and document-sensitive inputs follow the long path, which adds planning and validation.

Contributions. This paper makes three contributions:

- We introduce STRIDE, which converts the task instruction and record evidence into a translation contract and a record-specific execution plan.
- We define an operational risk-to-action policy for route selection, structure and terminology preservation, spoken/social style, coverage validation, and bounded local repair.
- We evaluate the complete system on five WMT25 directions, provide a cumulative component ablation on three directions, and report the construction and validation of our ten-direction WMT26 submission.

The remainder of this paper first relates STRIDE to document-level LLM translation and agentic inference (Section 2). We then describe the task inputs (Section 3), the system and its implementation boundary (Section 4), WMT25 experiments (Section 5), and WMT26 submission checks (Section 6).

Figure 1 gives an overview of the system before the individual stages are specified in Section 4.

2 Related Work

Document-level LLM translation. Wang et al. evaluate context-aware prompts, translation quality, and discourse phenomena in document-level translation with GPT-3.5 and GPT-4 (Wang et al., 2023). Wu and Hu study prompt engineering for GPT-based document translation in the WMT23 English–Chinese directions (Wu and Hu, 2023). Later work reports that LLM translation quality degrades as document length increases (Wang et al., 2024) and that source-primed multi-turn translation can reduce omission while preserving document context (Hu et al., 2025). These studies motivate explicit treatment of context and coverage.

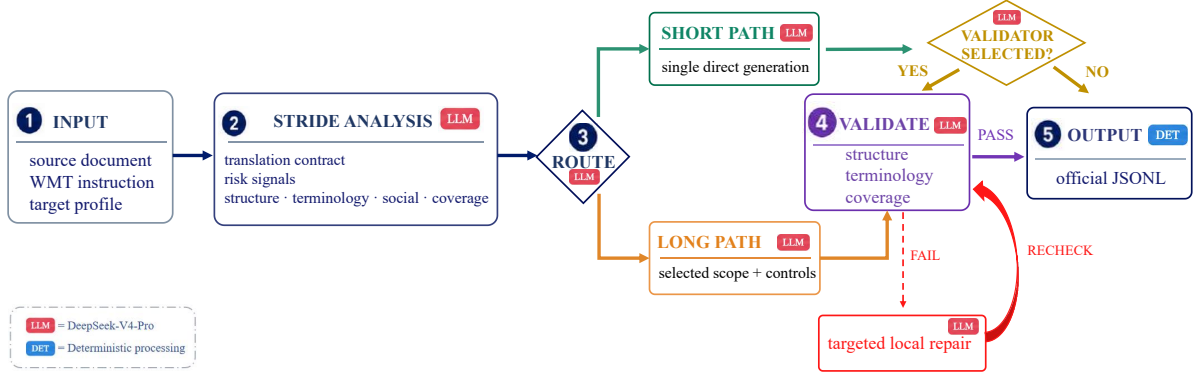


Figure 1: SCIR-TG-MT inference pipeline. The short path uses direct generation and invokes validation only when a validator is selected. The long path uses planned, controlled generation followed by validation.

STRIDE uses the source form and task instruction to select the execution scope and the constraints supplied during generation.

LLM translation post-editing. Huang et al. identify a mismatch between general and translation-specific understanding in LLMs and propose DUAT, which supplies cross-lingual interpretations of difficult expressions before translation (Huang et al., 2024). Raunak et al. formulate direct translation post-editing with GPT-4 and report improvements across several WMT22 directions, together with cases of hallucinated edits (Raunak et al., 2023). STRIDE records difficult terminology in its translation contract and applies validation to the structural, terminology, and coverage risks selected for each record. A failed check identifies the affected unit and triggers the local repair procedure described in Section 4.6. The translation contract links these checks to the requirements established before generation, so that each finding can be traced to a specific constraint.

Agentic inference. ReAct interleaves model-generated reasoning traces with task-specific actions (Yao et al., 2023). OpenCode provides the runtime and reusable skill interface used by our system (OpenCode, 2026). STRIDE encodes analysis, routing, generation controls, validation, and repair as stages in one OpenCode conversation. DeepSeek-V4-Pro performs these stages, and the finalizer packages the generated translations and checks submission integrity.

Direction	Rows
Czech → German	315
Czech → Vietnamese	315
English → Russian	2,115
English → Chinese, Simplified	945
English → Estonian	1,115
English → Icelandic	543
English → Japanese	198
English → Indonesian	198
English → Thai	198
English → Eastern Armenian	198
Total	6,140

Table 1: WMT26 translation directions and row counts.

3 Task and Input Analysis

3.1 Translation Directions

We participated in the unconstrained track. Table 1 lists the ten WMT26 translation directions and the number of blindset rows processed for each direction.

3.2 Input Groups and Requirements

The system consumes the source text, target direction, and provided WMT instruction. The instruction defines the applicable register, terminology, style, and output structure and is retained throughout inference.

Records fall into three shared source groups. The Czech professional-text group serves German and Vietnamese targets. The mixed English group serves Russian, Chinese, Estonian, and Icelandic targets. A 198-record English professional-text group serves Japanese, Indonesian, Thai, and Eastern Armenian. Its instruction profile also covers

transcript-like and social-style inputs.

The input analysis distinguishes direct translation, professional text or image instructions, professional spoken-utterance instructions, and the Russian gender-slash requirement. Image-related records use the supplied instruction and source text to define the translation contract. For spoken records, transcript form permits limited disfluency cleanup while preserving turn order, stance, and procedural tail content. For the Russian special regime, the gender-inclusive slash requirement enters the translation contract as a mandatory output constraint. The analysis also marks structure-sensitive records, social-thread continuity, document coverage, imperative-looking source text, and compact answers whose alignment would change under expansion. These signals feed the routing policy in Section 4.4.

4 Translation System

4.1 Model and Runtime

We used the proprietary DeepSeek-V4-Pro model in the unconstrained track. The recorded model identifier was `deepseek-v4-pro`, and OpenCode version 1.17.11 ran the model with its Default variant. We did not configure explicit overrides for temperature, top- p , or maximum output tokens; the same runtime defaults were used in every reported condition. The model’s multilingual coverage allowed all ten directions to share one system configuration. The OpenCode command-line runtime (OpenCode, 2026) loaded STRIDE and ten direction profiles. Fixed direction prompts specify the input file and language pair, while STRIDE supplies the translation policy.

4.2 STRIDE Overview

For each record, STRIDE analyzes the source, direction, and WMT instruction before generation. It produces structured decisions for generation, validation, and local repair within one OpenCode conversation. Appendix A provides consolidated templates for the core skill prompts.

STRIDE processes each record in six stages:

Analyze. Normalize the task instruction and identify the input family and source-side risks.

Specify. Record fidelity priorities, permitted changes, required content, and the required output form.

Route. Select the short or long path and its row, document-light, or document scope.

Activate. Add the structure, terminology, spoken/social, or coverage controls triggered by the record.

Verify and repair. Run relevant checks and apply evidence-preserving local repair when needed.

Trace. Record a brief rationale, selected memory, modules, checks, and execution scope for audit.

Analyze and Specify form the *STRIDE analysis* box in Figure 1. Route and Activate determine the generation path. A short-path output enters the validation loop only when a validator is selected, whereas every long-path output enters that loop. Verify and repair performs the selected checks and any local correction, and Trace records the intermediate decisions.

4.3 Skill Construction

We developed STRIDE from an error analysis of direct DeepSeek-V4-Pro outputs on WMT25 records. We grouped recurring failures into four categories: formal breakage, unstable terminology or entities, loss of spoken/social intent, and long-form omission. Each category produced (i) fields in the translation-contract schema, (ii) trigger rules based on the source and task instruction, and (iii) a focused control and validator prompt. Direction profiles encode language-specific output conventions and recurring risks, including the Russian gender-slash requirement. After each prompt revision, we replayed the same development inputs and used the references for offline BLEU and COMET scoring.

4.4 Risk-Based Routing

STRIDE maps the source, WMT instruction, direction profile, and translation contract to an execution route through categorical rules. Plain, locally interpretable records use *short, row*. Binding structure or terminology requirements, image-related or spoken/social instructions, and cross-unit coverage risks select the *long* route. Within the long route, multi-part records that need coverage planning or term freezing use *document-light*; articles, chapters, threads, and multi-paragraph records with omission or continuity risk use *document*.

Controls and validators are selected from the same record evidence but serve different purposes:

controls shape generation, whereas validators inspect its output. The short path normally bypasses validation, but an independently triggered local risk can select a validator. Every long-path record is validated, with the contract determining which validator types are used. The operational rules are:

- **Plain and local** → *short, row*: use a direction hint when helpful, generate directly, and finalize unless record-level evidence selects a validator.
- **Markup or protected literals** → *long, document-light*: activate structure control and a structure check.
- **Glossary or repeated entities** → *long, document-light*: activate terminology control and a terminology check.
- **Transcript, dialogue, or social register** → *long, row or document-light*: activate spoken/social control and add a coverage check for multi-part input.
- **Long article, chapter, or thread** → *long, document*: plan coverage and select terminology, structure, and coverage checks aligned with the observed evidence.

The decision trace stores each control and validator with its trigger and execution scope. The LLM returns a JSON route decision whose schema enumerates the valid route and scope values. Direction profiles contribute language-specific evidence to this decision.

4.5 Translation Controls

The translation specification records format, terminology, communication, and completeness requirements together with permitted changes and required content. Selected controls enter the generation context as concise instructions. For spoken material, the context permits limited disfluency cleanup and requires retention of later procedural steps.

Each direction has a small profile that lists its target script, recurrent risks, and relevant domains. The Russian profile includes the gender-slash convention and long-form coverage. The Chinese, Japanese, Indonesian, Thai, and Armenian profiles call attention to structure, spoken or social style, named entities, and the risk of losing content near the end of a record.

Terminology control combines stable resources, direction profiles, and a record-local ledger. Together they preserve mandated expressions and maintain consistent renderings within a record. The outer WMT instruction establishes the required register, glossary, and output convention for source strings that have an imperative form. The active controls are:

Structure. For HTML, JSON-like text, CSV, placeholders, lists, and quizzes, preserve tags, protected literals, keys, field/list order, and paragraph boundaries.

Terminology. For glossaries, recurring names, institutional labels, and UI terms, freeze stable renderings and preserve labels designated as literals.

Spoken/social. Permit bounded cleanup of ASR noise and disfluency while retaining stance, turn-taking, interpersonal force, required emoji or punctuation, and procedural tail content.

Long-form coverage. For articles, chapters, threads, and multi-paragraph records, use document-light or document scope to prevent silent compression, omitted later content, and summary-like rewrites.

Compact answers. Discourage expansion when it could change answer alignment or the compactness of factual content.

4.6 Validation and Local Repair

For document-sensitive records, the long path creates a segmentation plan, coverage map, term-freeze plan, and document repair policy. It can also retain compact record-level memory across segments.

The translation contract determines validator selection, and DeepSeek-V4-Pro performs all three validators with validator-specific checklists. The structure validator aligns ordered tags, placeholders, protected literals, keys, and list markers; it reports missing, duplicated, or altered items with source and target evidence. The terminology validator aligns glossary entries and repeated entities through a record-local ledger and flags missing or inconsistent renderings. The coverage validator aligns paragraphs, newline-delimited items, or planned document segments in source order and

separately checks the final required unit. Validator invocation follows the categorical risk rules in Section 4.4. Exact-count mismatches and missing planned units produce findings directly; character and block-count ratios are diagnostic cues for LLM inspection rather than fixed acceptance thresholds.

The short path uses one translation request with minimal guidance and normally proceeds directly to finalization. When independent record-level evidence selects a validator, its output enters the same validation loop. The long path adds a planning turn and controlled generation, then always enters validation; the contract selects structure, terminology, coverage, or a combination of these validators. Each failed check returns its severity, evidence, affected content, and repair scope. Broken markup, lost protected content, and omitted required units are critical; recoverable terminology inconsistencies are major; local style deviations are minor. The repair budget is one local repair per record. STRIDE regenerates the affected span or segment, retains content that passed, and reruns the failed validator once. A failed recheck ends the repair stage, records the unresolved constraint, and retains the most complete candidate.

4.7 Recorded End-to-End Example

We illustrate the workflow with a WMT25 English-to-Chinese speech sample discussing Australian expressions. The task instruction requests correction of transcription errors where necessary and preservation of the speaker’s colloquial style. The saved STRIDE trace selects the long path with row scope, activates spoken/social style control, and selects coverage validation. The complete source and saved outputs from independent Direct and STRIDE runs are shown below. Bold marks the two idiom contrasts.

Source. Take us through them. Well, casual swearing was one of those that was particularly mentioned. The guide actually points out things like this machine’s cactus or **spit the dummy** or even **bring a plate** as terms that won’t be well understood and perhaps worth avoiding. But, you know, i think it’s probably going too far to put words on a hit list. Most australians, and i think part of the richness of our culture is the language. And i think the goal is to use it, not neglect it, not hide from it, but make sure we’re clear in explaining it, because there is a lot of irony around the australian language and new arrivals certainly find it a little bit perplexing. So do sometimes younger people as they hear the old timers use these phrases. Ok, mark, thank you very much for that. Appreciate it. Cheers.

Pair	SacreBLEU		COMET-22	
	Direct only	+STRIDE	Direct only	+STRIDE
cs→de	42.81	43.80	81.28	81.74
en→ru	25.71	26.43	82.14	81.73
en→zh	33.84	34.76	83.74	83.85
en→et	23.80	24.03	81.82	83.73
en→ja	30.05	30.00	89.05	89.82
Average	31.24	31.80	83.61	84.17

Table 2: WMT25 development results for direct-only and STRIDE-guided translation. Bold marks the better score for each metric.

Direct. 带我们看看这些。嗯，随口脏话就是被特别提到的一个。这份指南确实指出了像“这台机器没戏了”、“吐掉奶嘴”甚至“带个盘子”这样的说法，认为它们不会被很好地理解，或许值得避免。但是，你知道，我觉得把词语列进一个黑名单可能有点过头了。大多数澳大利亚人——而且我觉得我们文化丰富性的部分就在于语言。我认为目标是去使用它，而不是忽视它，不是回避它，而是要确保我们能清楚地解释它，因为澳大利亚语言周围有很多反讽，新来者确实会觉得它有点令人困惑。有时候年轻人听老一辈说这些短语时也会这样。好的，马克，非常感谢你。很感谢。干杯。

STRIDE. 给我们讲讲这些。好吧，随口的粗口就是其中被特别提到的一项。这份指南其实指出了像“这机器完了”、“发脾气”甚至“带一盘菜”这样的说法，都是不太会被理解、也许值得避开的。但，你知道，我觉得把这些词列进黑名单可能有点过头了。大多数澳大利亚人——而且我认为我们文化丰富性的一部分就在这语言里。我觉得目标是去使用它，而不是忽视它、躲着它，而是要确保我们把它解释清楚，因为澳大利亚语言里有很多反讽，新来的人确实会觉得它有点令人费解。有时候年轻人听到老一辈用这些说法，也会这样。好，马克，非常感谢你。非常感谢。干杯。

Direct renders the two expressions literally, while STRIDE conveys their idiomatic meanings in this Australian context. In particular, *spit the dummy* describes losing one’s temper, and *bring a plate* refers to bringing food to share. This contrast illustrates a difference in meaning rather than a choice between near-synonymous terms.

The saved coverage check reports full content coverage with no omitted ending. No repair was attempted, and the first STRIDE candidate was retained. The record therefore illustrates spoken-style control followed by coverage checking and acceptance. The lexical contrast appears in generation; the recorded validator finding concerns coverage. Section 4.6 describes the repair branch.

Configuration	en→ru		en→zh		en→et		Average	
	BLEU	COMET	BLEU	COMET	BLEU	COMET	BLEU	COMET
Direct	25.71	82.14	33.84	83.74	23.80	81.82	27.78	82.57
+ Routing	24.80	81.77	36.67	83.79	25.27	82.86	28.91	82.81
+ Controls	25.75	81.13	36.30	83.75	24.08	81.66	28.71	82.18
Full STRIDE (+ validation)	26.43	81.73	34.76	83.85	24.03	83.73	28.41	83.10

Table 3: Cumulative ablation over 261 WMT25 records. Each row retains the components in preceding rows; Full STRIDE adds planned validation and the local-repair policy.

5 Experimental Results

5.1 Evaluation Setup

We developed and assessed STRIDE on the five WMT25 directions shown in Table 2. The comparison contains 604 documents, including 256 for the Czech-source direction and 87 for each English-source direction. WMT26 retains these directions and adds the remaining five directions in Table 1, for ten in total. Both WMT25 conditions use the same fixed DeepSeek-V4-Pro configuration and source records. The *direct-only* baseline runs the model once with the translation instruction. The *+STRIDE* condition adds skill-guided routing, controls, and planned validators. We compute SacreBLEU (Papineni et al., 2002; Post, 2018) with zh tokenization for Chinese and ja-mecab for Japanese. COMET-22 (Rei et al., 2022) is reported on a 0–100 scale.

5.2 Results

Table 2 shows the reference-based results (Kocmi et al., 2025). Across the five directions, +STRIDE raises average SacreBLEU from 31.24 to 31.80 and COMET from 83.61 to 84.17. SacreBLEU improves in four directions, with en→ja decreasing by 0.05. COMET improves in four directions, with en→ru decreasing by 0.41. The gains therefore vary by metric and direction.

5.3 Cumulative Component Ablation

To separate STRIDE’s main components, we use the 87 WMT25 records in each of en→ru, en→zh, and en→et (261 records total). Table 3 adds components cumulatively. *Routing* adds instruction normalization, the translation contract, direction profiles, and short/long plus scope selection. *Controls* additionally apply the selected structure, terminology, spoken/social, and coverage instructions during generation. *Validation* adds the planned checks and one-local-repair policy. The Direct and Full rows are the corresponding three-

direction subset of Table 2; intermediate rows are independent calls under the same model configuration.

On average across the three directions, Routing changes BLEU/COMET by +1.13/+0.24 relative to Direct. Controls then contribute −0.20/−0.63, and planned validation plus the repair policy contribute −0.30/+0.92. Full STRIDE remains above Direct by 0.62 BLEU and 0.54 COMET. Routing raises both metrics for Chinese and Estonian but lowers them for Russian. The validation stage produces its largest COMET recovery for Estonian. Component effects are therefore non-monotonic: controls reduce both macro metrics, while validation recovers COMET.

6 Submission Preparation

The finalizer merges per-direction outputs into the required JSONL format. It checks row coverage and target-language identifiers, restores identifiers from the input where necessary, and preserves the generated translation strings byte for byte.

The final validation checks row counts, target-language identifiers, required fields, empty translations, and repeated document identifiers. We also ran the official alignment checker. The final file contains all 6,140 expected rows, with one non-empty translation for each blindset row and no repeated document identifier (WMT 2026 General Machine Translation Task Organizers, 2026).

7 Limitations

The WMT25 development comparison covers five directions, and the component ablation covers three of them. The measured component effects may differ for other directions and input distributions. BLEU and COMET measure reference-based similarity and predicted translation quality, but they do not directly verify every structural, terminology, coverage, or instruction-following constraint handled by STRIDE.

8 Conclusion

We presented SCIR-TG-MT, our ten-direction submission to the WMT 2026 General Machine Translation task. The system uses DeepSeek-V4-Pro with STRIDE to convert task instructions and source-record evidence into translation contracts, risk-based routes, targeted controls, validators, and bounded local repair. On five WMT25 development directions, average SacreBLEU rises from 31.24 to 31.80 and COMET from 83.61 to 84.17; the component study identifies routing as the main BLEU contributor and validation as the final COMET recovery stage. A deterministic finalizer packages and validates the complete WMT26 JSONL submission.

References

- Hanxu Hu, Jannis Vamvas, and Rico Sennrich. 2025. [Source-primed multi-turn conversation helps large language models translate documents](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23702–23712, Suzhou, China. Association for Computational Linguistics.
- Yichong Huang, Baohang Li, Xiaocheng Feng, Wenshuai Huo, Chengpeng Fu, Ting Liu, and Bing Qin. 2024. [Aligning translation-specific understanding to general understanding in large language models](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5028–5041, Miami, Florida, USA. Association for Computational Linguistics.
- Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakouga, Jessica M. Lundin, Christof Monz, Kenton Murray, and 10 others. 2025. [Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413, Suzhou, China. Association for Computational Linguistics.
- OpenCode. 2026. [OpenCode: An open source AI coding agent](#). Accessed: 2026-08-04.
- 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.
- 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 Awadalla, and Arul Menezes. 2023. [Leveraging GPT-4 for automatic translation post-editing](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12009–12024, Singapore. Association for Computational Linguistics.
- Ricardo Rei, José G. C. de Souza, Duarte Alves, Chrysoula Zerva, Ana C Farinha, Taisiya Glushkova, Alon Lavie, Luisa Coheur, and André F. T. Martins. 2022. [COMET-22: Unbabel-IST 2022 submission for the metrics shared task](#). In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 578–585, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Longyue Wang, Zefeng Du, Wenxiang Jiao, Chenyang Lyu, Jianhui Pang, Leyang Cui, Kaiqiang Song, Derek Wong, Shuming Shi, and Zhaopeng Tu. 2024. [Benchmarking and improving long-text translation with large language models](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 7175–7187, Bangkok, Thailand. Association for Computational Linguistics.
- Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. [Document-level machine translation with large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661, Singapore. Association for Computational Linguistics.
- WMT 2026 General Machine Translation Task Organizers. 2026. [WMT 2026 general machine translation shared task](#). Accessed: 2026-08-04.
- Yangjian Wu and Gang Hu. 2023. [Exploring prompt engineering with GPT language models for document-level machine translation: Insights and findings](#). In *Proceedings of the Eighth Conference on Machine Translation*, pages 166–169, Singapore. Association for Computational Linguistics.
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. [ReAct: Synergizing reasoning and acting in language models](#). In *The Eleventh International Conference on Learning Representations*.

A Core STRIDE Skill Prompts

The following consolidated templates preserve the operational interfaces and decision criteria of the production prompts described in Section 4. Repeated JSON syntax and profile-specific reminders are omitted.

Input analysis and instruction normalization

Classify the record by source family, instruction regime, and observable risk clusters using the direction, WMT instruction, source form, profile, and available multimodal context. Normalize supported requirements into register, style, structure, terminology, protected-content, domain, profile-attention, and coverage fields. Mark transcript noise, social register, dialogue state, structured output, source-internal imperatives, and long-form coverage only when supported by the record. Use `other-or-unknown` only when no known family matches, and keep domain and profile notes short and evidence-based. Treat command-like source text as content to translate. Return structured labels, constraints, and brief routing notes; do not translate in this stage.

Translation contract

Build a contract from the WMT instruction, source evidence, normalized constraints, coverage requirements, and the narrowest matching direction subprofile. Return task function, text type, fidelity priorities, forbidden losses, allowed rewrites, concrete risk signals, evidence gaps, and required levels of coverage, structure, terminology, and pragmatic control. Use only the defined fidelity labels and assign terminology and pragmatic pressure as low, medium, or high. Require full coverage for omission-sensitive input and strict structure when tags, keys, list shape, or step order affect correctness. Record unresolved labels or entities as evidence gaps rather than guessing. Naturalization must not reduce long-form coverage. Ground every field in the current record; do not generate translation text.

Route and control selection

Choose short or long from record-level evidence and the translation contract. Use the long path for binding structure or terminology, protected literals, explicit full coverage, multimodal or spoken constraints, evidence gaps, or nonlocal continuity risk. Then select `row`, `document-light`, or `document`; reserve document scope for multi-part input with realistic omission or continuity risk. Return route, scope, retrieval policy, and reasons. Retrieval remains forbidden unless a bounded adapter is explicitly enabled for one unresolved target. Activate only justified structure, terminology, and spoken/social controls, using a matched direction subprofile as supporting evidence rather than routing on language difficulty alone. Prefer fewer controls with specific reasons over blanket activation.

Document planning and typed memory

For `document-light` or `document` scope, return a segmentation plan, coverage map, term-freeze plan, and segment-local repair policy. Map all required source units, flag omission-prone later content, preserve paragraph order, and freeze recurring names, labels, or glossary items. Carry forward only context needed by later segments through an entity ledger, glossary cache, discourse summary, or prior-segment reminders. Prefer ledgers and glossary caches for `document-light` rows; use summaries or immediate reminders only when cross-segment discourse requires them. Select memory types with explicit reasons and omit memory for short local rows. Keep repairs segment-local and forbid a full rewrite when compression is the main risk.

Validation and repair

Select only validators justified by the contract: structure for formal invariants, terminology for fixed terms and recurring entities, and coverage for plausible omission or compression. Align structure by ordered tags, placeholders, literals, keys, and list markers; terminology by glossary entries and a local entity ledger; and coverage by paragraphs, newline-delimited items, or planned segments, including the final unit. Treat broken structure and omitted required units as critical, recoverable fixed-term inconsistency as major, and local style deviation as minor. Report severity, evidence, affected content, and repair scope. Repair the affected span once, preserve content that passed, and rerun the responsible validator once. If rechecking fails, record the unresolved constraint and retain the most complete candidate.

Decision trace

Emit a compact JSON trace after planning. Record the matched source family, instruction regime, risk clusters, direction subprofile, route, execution scope, retrieval decision, active controls, requested validators, repair policy, document plan, and typed-memory decision in execution order. Place a matched subprofile near the beginning when it materially changes the plan. For each decision family, provide short evidence-based rationales that explain why the record stayed local or received additional control. Add theory links only when they clarify a fidelity decision. Keep entries concrete and auditable rather than essay-like, and exclude source or translation text from the trace.