

A Translator–Critic System for Chinese-to-English Subtitle Translation

Aaryan YekkiSETTY

Peloton Interactive

y.aaryan21@gmail.com

aaryan.y@onepeloton.com

Abstract

We introduce a text-only system that translates Chinese video subtitles into English using two open-weight models under a 20B-parameter budget. It is our submission to the constrained track of the WMT 2026 video subtitle translation shared task (WMT26 Organizers, 2026), whose rules set that budget. A 7B translation-specialized model generates English candidates, while a quantized 12B instruction-tuned model selects candidates and checks the selected translation for a small set of subtitle-specific accuracy errors. Confirmed errors are converted into focused correction instructions and returned to the translation model for guided retranslation. The system also preserves SubRip subtitle (SRT) indices and timestamps and applies deterministic name-consistency and line-layout processing.

We evaluate the main components through two internal studies scored by an independent large language model (LLM) judge. A synthetic-error experiment shows that the instruction-tuned model is better suited to the critic role than the translation model in our configuration, especially because it produces fewer false alarms on clean translations. An end-to-end ablation shows that contextual prompting provides the largest improvement over a context-free translator baseline; under a paired sign test it is the only component gain we can establish. The system placed third of eight in the official preliminary English evaluation at 74.705, 0.55 points behind second place. Those results are system-level and do not isolate individual components, but they localize the remaining gap to critical accuracy errors, the largest identified category of which is pronoun and referent confusion, which is what the critic was built to detect. This is consistent with the critic’s limited context window and single-model candidate pool. Reference translations were never released to participants, so no reference-based metric is computable on this test set.

1 Introduction

Video subtitle translation differs from conventional sentence-level machine translation (MT) in several important ways. Subtitle cues are short, often omit punctuation, and frequently contain only part of an utterance. A cue may therefore be difficult to translate correctly without information from neighboring cues, the series synopsis, or previously established character names.

These problems are especially visible in Chinese-to-English translation. Chinese is a pro-drop language in which subject and object pronouns are frequently omitted, particularly in conversational genres, while English translations often require them to be recovered explicitly (Wang et al., 2023); cross-sentence context is therefore necessary rather than merely useful (Voita et al., 2019). Pronoun gender may need to be inferred from context, terms of address may encode social rank (娘娘, roughly “Your Highness”; 皇上, “Your Majesty”) and rank-indexed self-reference (本宫, used by a ranking consort; 朕, the imperial “we”), and short expressions can have different interpretations depending on the speaker and scene. Character names must also remain consistent across an episode.

The provided test set makes this contextual dependence clear. It contains 41,837 cues across 100 SubRip subtitle (SRT) files. The median cue contains six Chinese characters, and 42.5% of cues contain five characters or fewer. None of the cues ends in terminal punctuation. As a result, punctuation-based sentence reconstruction is not available, and many cues must be interpreted as fragments of a longer exchange.

Providing context to a large language model is straightforward: neighboring cues and metadata can be included in the prompt. However, this introduces a second problem. The model may translate content from neighboring cues or merge sur-

rounding information into the target line. We refer to this behavior as *context leakage*. The system must therefore use context for interpretation while ensuring that only the target cue is translated.

Our system addresses this problem with a translator–critic pipeline. A translation-specialized model produces candidate translations using the source cue, neighboring cues, and series metadata. A separate instruction-tuned model selects among the candidates and audits the result using eight focused accuracy checks. When an error is confirmed, the translation model receives a short instruction naming the problem and produces new candidates.

The system is implemented with three main design choices, which together form our contribution:

1. **Context is presented as read-only background information.** Only the current cue appears in the translation field.
2. **Translation and criticism use different model families.** The translation-specialized model generates source-based translations, while the instruction-tuned model performs selection and post-editing.
3. **Auditing is decomposed by error type.** Each critic call checks one narrowly defined defect, followed by a second verification step before repair.

In this paper we focus on the submitted pipeline, its implementation under the constraints described above, controlled internal comparisons of its main components, and the official preliminary task results (§7).¹

2 Related Work

Document-level and context-aware MT. The observation that sentence isolation is the wrong unit for translation predates LLM pipelines. Tiedemann and Scherrer (2017) extend the translation window across sentence boundaries, and Voita et al. (2019) show that a translation can be correct in isolation and wrong in context, precisely the failure mode that dominates short subtitle cues. Pronoun recovery under Chinese pro-drop has been treated as its own problem (Hardmeier, 2015; Wang et al., 2023). Our context construction (§4.3)

is a deliberately minimal instance of this line of work: a fixed symmetric window of neighboring source cues plus album-level metadata, with no learned context selection.

Recent work replaces a fixed window with an agent that accumulates document state. Wang et al. (2025) maintain multi-level memory across a document for online document-level translation, and Ren et al. (2026) internalize document context into model parameters dynamically rather than re-reading it in the prompt. Both maintain learned or parametric document state where we keep a fixed window plus a name memory, and subtitle cues give them less to work with than prose: no cue in the corpus ends in terminal punctuation, so there is no reliable sentence or paragraph segmentation (§3). A document-state agent operating over a whole episode is the most promising direction we did not take.

Multi-agent and MQM-guided refinement. Assigning generation and criticism to different agents is an established pattern. Madaan et al. (2023) keep both roles in one model; Wu et al. (2025) distribute translation, evaluation, and revision across a collaborating group for ultra-long literary text. Error-typed feedback in particular has been shown to be more useful than a scalar score: Fernandes et al. (2023) elicit MQM-style error spans from an LLM rather than a single quality number, following the human MQM protocol of Freitag et al. (2021). Our critic is a restricted version of this idea. It does not produce a free-form MQM annotation; it answers eight fixed yes/no questions (§4.5), each in its own call, and a confirmed error is converted into a correction hint rather than an edit. The distinguishing choice is not the use of typed feedback but where repair happens: the translator generates new source-conditioned candidates and the critic selects among them, making limited English-side edits at most (§4.6), which resembles source-aware automatic post-editing (Chatterjee et al., 2020) rather than free rewriting.

Quality-aware decoding. Our candidate-and-selection stage (§4.4) occupies the same position in the pipeline as minimum Bayes risk decoding with neural metrics (Freitag et al., 2022) and quality-aware decoding with learned rerankers (Fernandes et al., 2022), which select from a candidate pool using a trained quality model. Our selector is an instruction-tuned LLM given the source, the can-

¹The system described here is our constrained-track zh→en shared-task submission, entered under the team name AuditTranslate.

Property	Value
Files / cues	100 / 41,837
Cues per file: min / median / max	41 / 279.5 / 2,188
Chinese chars per cue: median / p90	6 / 11
Cues with at most five characters	42.5%
Cues ending in terminal punct.	0
Cues with sentence punctuation	4
Single-line cues	99.67%
Cue duration: median / p90	1.28 s / 2.36 s
Episodes with a synopsis	100 / 100
Distinct series	45

Table 1: Summary of the released Chinese subtitle corpus. Every figure is recomputed from the released files.

didates and context, so it is a prompted reranker rather than a learned one. We did not compare against these methods.

3 Task and Data

The released test set contains 100 SRT files from 45 series. Each file includes subtitle cues and album-level metadata, including a title and synopsis. We use the Chinese subtitle text, series title, synopsis, and neighboring subtitle cues. We do not use video frames or audio. Table 1 summarizes the corpus.

The corpus covers several genres. Of the 100 episodes, 43 are xianxia or cultivation fantasy, 24 are palace or costume dramas, 16 are reality, variety, or talk programs, nine are modern urban dramas, and eight belong to other categories. Genre labels were assigned by hand over all 45 series from the released `album_title` and `album_desc` fields.

Genre affects the translation problem. Period and fantasy dramas contain rank-sensitive forms of address and specialized terminology, while modern and variety programs rely more heavily on conversational phrasing and real personal names. We therefore avoid assuming that one terminology mechanism will benefit every file equally.

Development procedure. No task-specific reference translations were available. During development, we inspected six files and used qualitative model reports to identify recurring errors and refine the pipeline. These files were not used for the quantitative studies reported below.

The critic experiment samples examples from 84 non-development files. The end-to-end ablation uses 14 non-development files selected to cover the main genre groups. We report the two studies separately because they answer different questions: the first evaluates critic behavior on

controlled errors, while the second evaluates complete translation outputs.

4 System Description

This section describes the submitted system. We first state the two model checkpoints and the parameter budget they have to fit into (§4.1), then follow a single cue through the pipeline in the order the stages run (§4.2–§4.7), and close with how the models are served (§4.8).

4.1 Models and parameter budget

The system uses two open-weight checkpoints: Hy-MT2-7B (Zheng et al., 2026; Tencent Hunyuan, 2026), a translation-specialized model; and Gemma-4-12B-it (Google, 2026), an instruction-tuned general-purpose model, used in its 4-bit integer quantization-aware training (INT4 QAT) release.

Both models are used at their released revisions without fine-tuning, adapters, tokenizer modifications, or chat-template changes. Following the checkpoints’ tied-embedding configuration, the total parameter count is approximately 19.47B: 7.505B for Hy-MT2 and 11.960B for Gemma. This remains under the constrained track’s 20B limit.

Hy-MT2 generates all initial and hint-guided translations. Gemma selects and post-edits translation candidates, performs rule-specific audits, and consolidates repairs. Gemma is not used as a standalone source-to-target translation system.

We use the quantized Gemma release because it runs within the 24 GB of memory available on our graphics processing unit (GPU) instances. Each model is served on a separate endpoint with one GPU.

4.2 Pipeline overview

Each SRT file passes through seven stages:

1. parse the SRT and construct contextual information;
2. generate up to three translation candidates;
3. select the best candidate;
4. audit the selected translation;
5. verify detected errors;
6. generate and select repairs when needed;
7. apply deterministic consistency and layout processing.

Cues can be translated concurrently because source-side neighboring cues are available before inference. Whole-file consistency processing is applied after cue-level translation. All prompts are reproduced in Appendix A.

4.3 Context construction

For cue i , the system uses up to two preceding and two following Chinese cues, together with the series title and synopsis. Neighboring cues are provided as background information and are explicitly marked as context that should not be translated.

Only the current source cue is placed in the translation field. This separation is intended to let the model resolve omitted subjects, pronouns, sentence mood, and short ambiguous expressions without encouraging it to translate surrounding lines.

4.4 Candidate generation

Hy-MT2 generates up to three candidates for each cue. The first draw uses temperature 0.7. The second and third draws use higher temperatures, 0.95 and 1.0, with top- $p=0.95$ and top- $k=40$.

The higher temperatures are used because short subtitle cues often produce identical outputs under repeated low-temperature sampling. Candidates are deduplicated exactly. In the production run, 43.97% of cues still produced only one distinct candidate.

If multiple distinct candidates are available, Gemma receives the current Chinese cue, neighboring Chinese context, series metadata, the candidate translations, and available character-name information. It returns a selected or minimally edited English line. The selection prompt emphasizes adequacy, name consistency, pronoun gender, register, brevity, and avoidance of context leakage. If only one distinct candidate is available, the selection stage is skipped.

4.5 Rule-specific audit

After selection, Gemma checks the translation for eight error types, listed in Table 2. Each error type is evaluated in a separate call. The categories are a subtitle-specific decomposition of primarily accuracy-related errors, inspired by Multidimensional Quality Metrics (MQM) (Freitag et al., 2021) rather than taken from it: several of the rules are our own task-specific categories with no one-to-one MQM equivalent. Each maps one-to-one onto a repair hint, so that the rule that fires is also

Rule	Purpose
meaning_preserved	Detects hallucination, omission, polarity errors, and major semantic distortion.
pronoun_gender_resolved	Checks whether an English pronoun or subject matches the contextual referent.
no_context_leak	Detects content copied or inferred from neighboring cues but absent from the target cue.
honorific_register_consistent	Checks titles, ranked address, and socially marked self-reference.
sentence_mood_match	Checks question, statement, and command distinctions.
direction_agent_correct	Detects direction, agent, patient, give/take, or come/go reversals.
person_attribution_correct	Checks first-, second-, and third-person attribution and terms of address.
no_invented_name	Detects proper names introduced where the source contains only a title, role, or pronoun.

Table 2: Error categories used by the critic. The exact question asked for each is reproduced verbatim in Table 7, and the repair hint each produces in §A.6.

what the translator is told when the line goes back for retranslation.

The audit focuses on adequacy-related errors rather than general fluency. Hy-MT2 normally produces fluent English, while broad style instructions can encourage unnecessary rewriting.

Each audit call returns a Boolean decision and a short explanation. A positive result is not immediately used for repair. Instead, a second Gemma call independently rechecks the same rule. This verification prompt defaults to “no defect” and confirms a problem only when it can identify concrete offending text.

This second stage is intended to reduce unnecessary repairs. A missed error leaves the translator’s original output unchanged, while a false positive may trigger a change to a line that was already acceptable.

4.6 Hint-guided repair

When one or more errors are confirmed, each rule is mapped to a short correction instruction. For example, a context-leak error produces an instruction to translate only the current cue; a pronoun error asks the translator to use the referent supported by context; and a mood error asks it to preserve whether the cue is a question, statement, or command.

These instructions are added to the Hy-MT2 translation prompt. The translator then generates two new candidates. Gemma considers the orig-

inal and new candidates and returns one consolidated line intended to address all confirmed defects.

The repaired line is checked once more against the confirmed rules. Remaining issues are recorded rather than repeatedly applying unconstrained repair.

4.7 Deterministic output processing

After cue-level translation, code performs three final operations: a whole-file character-name consistency pass; an English ordinal-title normalization, such as changing “Prince Ninth” to “Ninth Prince”; and line wrapping with a soft limit of 42 characters per line and a hard limit of 84 characters per cue.

The system then reconstructs each SRT file while preserving the original cue indices and timestamps. All 41,837 cues were processed without missing subtitle entries, and the output files retained the original indices and timing fields.

4.8 Serving and inference volume

Both models are served through vLLM (Kwon et al., 2023) on Amazon SageMaker, with one 24 GB GPU per endpoint.

Over the production run Hy-MT2 was called approximately 130,000 times and Gemma approximately 366,000 times, the latter dominated by the eight audit calls per cue. Appendix B gives the instance types and the exact call breakdown.

The high-volume audit calls use deterministic decoding with native reasoning disabled. Reasoning is enabled only for the lower-volume verification and repair stages.

5 Critic Configuration

Two choices in the system above need evidence: that the critic should be a different model from the translator, and that the audit should be decomposed into one call per rule and then gated by a skeptical verification pass. This section measures both on a controlled set of Chinese-to-English cases with known labels, holding the task text fixed within each configuration so that model identity is the only variable in the 7B-versus-12B comparisons.

5.1 Controlled evaluation set

The evaluation set contains 600 Chinese-to-English cases: 400 translations with injected errors, 50 for each audit rule; and 200 unmodified controls.

Configuration	Recall	Clean flag rate
7B, one combined call	49.5%	58.5%
7B, one call per rule	81.8%	77.5%
7B, per rule + verification	13.5%	2.5%
12B, one combined call	88.5%	12.0%
12B, combined + verification	84.0%	2.5%
12B, one call per rule	98.0%	12.5%
12B, per rule + verif.*	91.5%	3.5%

Table 3: Detection recall and clean-control flag rate on the controlled critic set. Each row is one critic configuration run over the same 600 labeled cases; within a configuration both models receive identical task text, while prompt structure differs across configurations as described in §5. Recall is rule-level over the 400 injected errors: the correct rule had to fire. For the flag rate, lower is better. *is the submitted configuration.

Injected errors are created through deterministic English-side transformations. Examples include flipping pronoun gender, inserting or removing negation, reversing an agent and patient, changing sentence mood, flattening an honorific, or inserting content from a neighboring cue.

Eligibility filters are applied before each transformation. For example, a pronoun is modified only when the Chinese source provides evidence about the referent, and an agent–patient reversal is used only when the source supports the original roles. Transformations that would make the English ungrammatical are excluded.

The resulting set tests detection of controlled errors from the system’s target taxonomy. It does not represent the full distribution of naturally occurring translation errors.

5.2 Configurations

We compare one call containing the complete rubric against one call per rule, with and without skeptical verification. Seven of the eight resulting configurations were run; the 7B combined rubric with verification was not. Both models receive the same task text at temperature zero. We report *recall* over the 400 injected errors, requiring the intended rule to fire, and the *clean-control flag rate* over the 200 unmodified examples. Results are given in Table 3.

5.3 Results

In this setup, the instruction-tuned 12B is better suited to criticism than the translation-specialized 7B. Without verification, the per-rule 12B configuration detects 98.0% of injected errors and flags 12.5% of clean controls. The corresponding 7B

setup detects 81.8% of injected errors but flags 77.5% of clean controls.

The main difference is therefore not only error detection. The 12B is substantially more selective on clean inputs. This matters because most production translations are not repaired, and unnecessary intervention can degrade an acceptable subtitle.

For the 12B, decomposing the rubric into separate calls increases observed recall from 88.5% to 98.0%, while the clean-control flag rate remains similar (12.0%→12.5%). Adding skeptical verification reduces recall to 91.5% but lowers the clean-control flag rate from 12.5% to 3.5%. We use this configuration in the submitted system. Note that verification is not a universal precision fix: applied to the 7B it collapses recall from 81.8% to 13.5%, because under a skeptical prompt the specialist clears most of its own initial flags.

The lowest-recall rule in the final setup is `honorific_register_consistent`, at 70%. Honorifics are difficult because their correct rendering depends on genre, speaker status, and local discourse context.

These results should be interpreted as controlled detection measurements. The errors are synthetic, and the clean controls are sampled from system output rather than independently human-verified references.

6 End-to-End Evaluation

The second experiment asks whether the complete pipeline translates better than the translation model alone. We build four cumulative configurations over the same held-out cues, score them with an independent judge model, and then look separately at the small subset of cues the critic actually edits.

6.1 Systems

We compare four cumulative configurations:

- S1 Translator only.** One Hy-MT2 draw without neighboring or synopsis context.
- S2 + Context.** One Hy-MT2 draw with neighboring cues and series metadata.
- S3 + Candidate generation and selection.** Three Hy-MT2 draws followed by Gemma selection.
- S4 Full system.** S3 followed by audit, verification, guided repair, name consistency, and layout processing.

S2 and the first draw of S3 use the same translation prompt, allowing S2→S3 to measure the combined effect of additional candidates and external selection.

The four outputs were generated for 4,563 cues from 14 non-development files. A common subset of 612 cues was scored across all four arms. The critic-edited subset was also examined separately because the critic changes only a small fraction of all cues.

6.2 Judge

Neither pipeline model is used as the external evaluator: Hy-MT2 generates the candidates and Gemma performs selection and repair, so scoring with either would be a same-model evaluation, in which a model tends to prefer its own output (Wataoka et al., 2024). We therefore use a separate, larger model, `gpt-5.6-luna`, with high reasoning effort; because the judge plays no part in generating the submission, it is not bound by the shared task’s parameter budget. Using a strong language model as an MT evaluator is established practice (Kocmi and Federmann, 2023), including for fine-grained MQM-style error annotation (Fernandes et al., 2023).

Each evaluation item contains the Chinese target cue, neighboring Chinese cues, and two anonymized English renderings. System identities are not shown. Candidate order is reversed in a second evaluation so that each pair is judged in both presentation orders. Identical translations are collapsed to avoid presenting duplicated outputs as independent agreement.

The judge assigns three 0–100 scores (adequacy, readability, and overall quality) and up to two MQM-style errors per candidate (Zheng et al., 2023). Following Freitag et al. (2021), minor and major errors receive weights of 1 and 5; we additionally define a task-specific critical weight of 10 for meaning inversions.

This is a single-model evaluation. We use blinding and order counterbalancing to reduce obvious presentation bias, but the resulting values remain model-based estimates rather than human judgments.

6.3 Main results

The largest observed improvement comes from adding series metadata and neighboring source cues (Table 4). The overall score increases by 2.32 points from S1 to S2, while the MQM penalty falls

System	Overall	Adeq.	Read.	MQM↓
S1 translator only	87.18	87.47	92.93	1.63
S2 + context	89.50	90.11	93.52	1.27
S3 + cand. & select	90.44	90.51	94.78	1.14
S4 full system	90.57	90.65	94.80	1.10

Table 4: Mean scores from the independent LLM judge over the common evaluated sample of 612 cues. The MQM error penalty column runs the other way: lower is better. Paired transitions are reported separately in Table 5.

Transition	Δ	95% CI	W/T/L	p
S1→S2	+2.32	[+0.94, +3.68]	189/304/119	<0.001
S2→S3	+0.93	[+0.06, +1.87]	128/349/135	0.71
S3→S4	+0.13	[-0.20, +0.41]	8/600/4	0.39
S1→S4	+3.39	[+2.04, +4.80]	223/227/162	0.002

Table 5: Paired analysis of each transition, measured cue by cue between the indicated system pairs over the 612 cues judged in every arm. Δ is the mean change in overall score. Confidence intervals are percentile bootstrap intervals resampled over cues (4,000 resamples); the cue is the sampling unit, so repeated judgments of one cue are averaged before resampling rather than treated as independent. W/T/L counts cues that improve, tie, and worsen, and p is an exact two-sided sign test on those counts with ties excluded. The sign test ignores effect size and therefore asks a separate question (whether *more* cues improve than worsen), and the two questions disagree for S2→S3.

from 1.63 to 1.27. This supports the use of contextual prompting for short, punctuation-free subtitle fragments.

Candidate generation and selection add a further 0.93 points, and the complete system improves by 3.39 points over the context-free one-pass baseline and by 1.07 points over the context-aware one-pass baseline. As §6.4 shows, only the first and last of these differences survive a paired test.

After reweighting the genre distribution to match all 100 files, the overall scores are 87.82, 89.78, 90.62, and 90.70 for S1 through S4. The corresponding full-system improvements are therefore 2.88 points over S1 and 0.92 points over S2.

These differences are point estimates from one evaluator and should not be interpreted as calibrated human-quality units.

6.4 Statistical reliability

Table 5 reports paired bootstrap intervals and sign tests for each transition. The four transitions do not stand on equal evidence, and the ordering in Table 4 should not be read as four established gains.

S1→S2 (context) is the one robust component

result: the interval excludes zero and 189 cues improve against 119 that worsen ($p < 0.001$). S1→S4 (the full system) is likewise robust ($p = 0.002$), and the evidence attributes most of it to context. S2→S3 (candidates and selection) is not established: the interval barely excludes zero, but 128 cues improve while 135 worsen ($p = 0.71$), so the positive mean comes from effect size on a minority of cues rather than from broad improvement. That transition is therefore suggestive rather than demonstrated. S3→S4 (the critic) is indistinguishable from noise in aggregate ($p = 0.39$), which is the expected consequence of a mechanism that touches about 2% of cues rather than evidence that it does nothing; §6.5 analyzes the treated cues directly.

These intervals quantify sampling variation over cues for one evaluator, not evaluator bias, and are computed on 612 cues from 14 files.

6.5 Effect of the critic

The aggregate S3→S4 difference is small because the critic is deliberately conservative. It changes approximately 2% of cues.

On the 94 cues where the critic changes the translation, 70 receive a higher score, two are tied, and 22 receive a lower score. Thus, 74.5% of all changed cues improve according to the external judge, or 76.1% when ties are excluded. The mean overall change on this subset is +14.8 points.

The gain is concentrated in adequacy (+17.2) rather than readability (+0.1). This is consistent with the critic’s design. Its rules focus on meaning, attribution, pronouns, context leakage, and sentence mood rather than stylistic rewriting.

Two examples illustrate the type of correction. 还喃喃自语 (“still muttering to oneself”) was initially translated as “muttering to *herself*” and changed to “muttering to *himself*,” based on a male referent in the preceding context. 和你妹的事 (“the matter concerning your sister”) was initially rendered as “Nothing to do with your sister” and changed to “something about your sister,” correcting a polarity error.

The critic is therefore best viewed as a targeted correction mechanism rather than the main source of the system’s average improvement. Most of the aggregate gain comes from contextual translation and candidate selection.

7 Official Preliminary Results

Official preliminary results became available after the internal evaluations (WMT26 Video Subtitle Translation Organizers, 2026), and they supply the independent evaluator our internal study lacked: a different model family, a different rubric, and all 100 files rather than 14. They assess the submitted system as a whole and do not isolate the contribution of any individual component, but they localize where its remaining headroom is.

7.1 Protocol and standing

The official evaluation scores each (team, language, video) unit with gemini-3.1-pro-preview under a language-specific translation-quality prompt, three independent runs per video with the median taken, followed by a human verification stage covering 93.4% of units across all five language pairs and 82.4% of the English units. Dimension scores are recomputed in code from itemized defect penalties on a 100-point deduction scale: Accuracy/Fidelity 40, Fluency/Readability 35, Cultural/Stylistic 20, and Subtitle Text Conventions 5.

Our system scored **74.705 in English, placing third** of eight teams in the only direction we entered (Table 6). The second-, third-, and fourth-ranked systems are separated by adjacent gaps of 0.55 points, spanning 1.10 points overall. The gap to ReopenAI is of a different order: −15.49 points, with ReopenAI scoring higher on 92 of the 100 videos.

7.2 The gap is one dimension, at one severity, on one phenomenon

Attributing the difference to ReopenAI by dimension gives +12.95 Accuracy/Fidelity, +1.57 Cultural/Stylistic, +1.33 Fluency/Readability and −0.19 Conventions, summing to 15.66 against a 15.49-point score gap. Roughly five sixths of it is Accuracy/Fidelity. Conventions is a five-point band in which every leading system is near ceiling, so the deterministic layout and name processing of §4.7 had little room to matter either way.

Severity localizes the gap further. Counting Accuracy defects by penalty weight, our minor-error count is comparable to the winner’s (153 defects at penalty 1 versus their 161), while critical severity-5 defects differ by more than an order of magnitude: 233 versus 19, or 2.33 versus 0.19 per video. The system is not broadly less accurate. It fails

Team (English)	Score	Mean points lost per video			
		Acc. /40	Flu. /35	Cult. /20	Conv. /5
ReopenAI	90.195	3.80	3.74	1.57	0.64
LensLinguals	75.255	15.43	4.80	3.54	0.55
AuditTranslate	74.705	16.75	5.07	3.14	0.45
laniqo	74.155	17.22	4.80	3.52	0.32
rupiah	70.895	18.89	5.74	4.41	0.65
BVSLP	60.480	n/a	n/a	n/a	n/a
paramanandaAI	55.645	n/a	n/a	n/a	n/a
SubVision	49.775	n/a	n/a	n/a	n/a

Table 6: Official preliminary English-track scores (WMT26 Video Subtitle Translation Organizers, 2026), with rubric-dimension attribution for the five leading systems. Lower is better in the four right-hand columns. Dimension means are computed from the itemized defect penalties, clipped at each rubric maximum, over the units with a parseable itemized assessment (98–100 per team), while the headline score averages all 100; the two are therefore close but not additive, so the dimension columns attribute where points are lost rather than exactly decomposing the score gap.

badly on a small number of cues, which is also consistent with the score distribution: our median (75.75) is essentially LensLinguals’ (76.0), but our 10th percentile is 55.5 against their 64.0, and 14 of our videos score below 60 against their 6. What separates us from second place is the lower tail, not the typical video.

Within those 233 severity-5 Accuracy defects, the largest single category is pronoun, referent and person confusion: 52 defects, 22.3% of the total, costing 2.60 penalty points per video. This is a heuristic count, obtained by keyword rules over the judge’s free-text descriptions and subject to errors in both directions; we did not hand-validate the 52 cases, and a 36.5% residual resists the classification altogether.

Cross-cue reference resolution is therefore where our remaining headroom is concentrated, and it is also what the critic’s decomposed checks were written for: `pronoun_gender_resolved`, `direction_agent_correct` and `person_attribution_correct` are three of the eight rules in Table 2. That these errors reach the final output is consistent with the audit’s evidence being too local: the critic sees the same ± 2 cue window as the translator, so a referent established twenty cues earlier lies outside it, whereas ReopenAI supplied the complete subtitle sequence to both generation and selection. The candidate pool may limit it the same way, since all three candidates were drawn from Hy-MT2 alone and a

selector cannot choose a rendering no candidate contains. We cannot separate these explanations from weak audit instructions, over-skeptical verification or repair failure, and the comparison with ReopenAI varies several factors at once; an episode-level window and a two-model candidate pool are the changes we would test first.

One further defect we would not have found without the official records: on cues of three Chinese characters or fewer, output is occasionally displaced by one cue, carrying the following line’s content. Measured directly in our submitted files, this affects 83 cues across 46 videos. The judge names the failure independently on a partly overlapping set, charging 33 defects across 27 videos as a translation shift, worth 1.25 penalty points per video. A source-to-target length-ratio check could flag some such cases for review, though it would not reliably catch all of them.

8 Discussion

The two experiments above measure what the pipeline buys. This section explains why its main mechanisms are built the way they are, covering read-only context, criticism separated from translation, and repair routed back through the translator, and what that design costs at inference time.

Why context is useful. A large fraction of the corpus consists of very short fragments. In such cases, the current cue may not explicitly identify the speaker, the referent of an omitted subject, pronoun gender, whether the line continues a question or command, or the intended interpretation of an honorific or title.

The context-aware configuration improves adequacy more than readability, suggesting that neighboring information mainly helps determine what the cue means rather than how fluent the English sounds.

However, contextual prompting requires strict source boundaries. Neighboring cues are therefore placed in background fields, while only the current cue is placed in the translation field (§4.3). The explicit `no_context_leak` rule provides an additional check.

Why criticism is separated from translation. Self-refinement (Madaan et al., 2023) asks one model to critique and revise its own output, and is the closest single-model analogue of what our selection, audit and repair stages do. We separate the

roles instead. Hy-MT2 is specialized for translation, while Gemma is better at following structured classification and verification instructions in our controlled experiment. The two-model arrangement uses these capabilities differently: Hy-MT2 generates source-grounded translation candidates; Gemma compares candidates and detects specific defects; and confirmed problems are returned to Hy-MT2 as focused guidance.

This arrangement also avoids asking the translation model to be the only evaluator of its own output (Wataoka et al., 2024). It does not eliminate evaluator bias, but it provides an independent model family for selection and auditing.

Why repair is routed back through the translator. Early development tests showed that allowing an editor to freely rewrite a line could introduce unsupported content. We therefore use the critic primarily to identify and describe errors. New source-based candidates are then generated by the translation model with a correction hint, so the mechanism functionally resembles source-aware automatic post-editing (Chatterjee et al., 2020) rather than free rewriting.

The final selection stage can still make limited English-side edits, so the separation is functional rather than mechanically guaranteed. A stricter future version could require the selector to return an exact candidate index.

Cost and intervention rate. The audit is computationally expensive because it applies eight checks to every cue, while only a small fraction of cues are repaired. This design prioritizes focused detection and a low false-alarm rate over inference efficiency.

The controlled experiment indicates that a single combined audit call with verification is a cheaper alternative, although it has lower synthetic-error recall than the per-rule configuration. A useful future direction is to compare these configurations directly on final translation quality and GPU cost.

In light of §7 we would now allocate this cost differently. The eight-per-cue audit stage was the largest single consumer of inference in the pipeline, and each call reasoned over a ± 2 cue window too narrow to see the cross-cue referents that dominate our remaining accuracy gap. Spending part of the same budget on full-episode context would let the rules in Table 2 operate on the evidence they need: size the evidence available to each stage before scaling the number of checks it performs. The lim-

itations of the system and of the evidence above are collected in the Limitations section.

9 Conclusion

We presented a text-only Chinese-to-English subtitle translation system built from two open-weight models under a 20B parameter limit. A translation-specialized 7B model generates candidates, while a quantized instruction-tuned 12B model performs selection, targeted auditing, and limited post-editing. Confirmed errors are routed back to the translation model through focused retranslation hints.

The internal ablation identifies contextual prompting as the largest source of improvement. Adding series metadata and neighboring cues improves the external judge’s overall score by 2.32 points over a context-free translator, and this is the one component gain that a paired sign test establishes. Candidate generation and selection add a further 0.93 points on average, but roughly as many cues worsen as improve, so we describe that transition as suggestive rather than demonstrated. The complete system scores 3.39 points above the context-free baseline on the common evaluated sample.

The official preliminary evaluation placed the system third of eight in English at 74.705, 0.55 points behind second place and 15.49 behind the winner. It assesses the system as a whole rather than isolating components, but it localizes the remaining gap: most of it is Accuracy/Fidelity, most of that is critical severity, and the largest category we identify within it is pronoun and referent confusion, which is what three of the eight audit rules were written to catch.

The system’s main practical lesson is therefore about what a critic needs rather than whether to have one. Separating generation from criticism supported targeted repairs, but it required substantial inference (roughly 366,000 critic calls, dominated by eight audit calls per cue), and it did not eliminate harmful edits: 22 of the 94 cues the critic changed scored lower. A critic can confirm only errors it has the evidence to see, and can choose only among renderings some candidate contains. Ours was given a ± 2 cue window and a candidate pool from one model; full-episode context and candidates from both models are the changes we would test first.

Limitations

LLM-based quality evaluation. Our component ablations (§6) use a single independent LLM judge on 612 cues from 14 files, with no bilingual human annotation. Blinding and presentation-order reversal reduce some sources of bias but do not replace human evaluation, and a 14-file sample cannot resolve per-component differences of the size we report. The official evaluation (§7) covers all 100 files, but it is also LLM-based and reference-free, includes no significance testing, and its human stage verified automatic assessments rather than annotating independently. It is also system-level, so it cannot arbitrate between components. All per-component attributions in this paper therefore rest on a single judge and are provisional.

Synthetic critic errors. The critic study uses controlled English-side perturbations. These provide known target error categories, but they are more regular than naturally occurring translation errors. The clean controls are also sampled from system output rather than human-verified references.

Text-only input. The system does not use video or audio. Visual and acoustic information could help resolve speaker identity, gender, sentence mood, and scene boundaries. The current system relies only on subtitle text and metadata.

Genre prompting. Some selection and repair prompts used a historical-drama editor role across all files. This matches much of the corpus but is not appropriate for every modern or variety program. A genre-neutral or metadata-conditioned editor prompt would be preferable.

Ultra-short cues. Very short cues remain difficult. Of the 125 cues containing only 是 (“yes”), approximately one quarter were expanded into six or more English words, often because the model used neighboring context too aggressively. A source-to-target length guard or duration-aware shortening stage could reduce such expansions.

Subtitle timing constraints. The system preserves timestamps and applies line-length limits (Karakanta et al., 2020), but it does not optimize translations directly for reading speed based on cue duration. Duration-aware generation or shortening would provide stronger subtitle-specific control.

No reference-based evaluation. Official task scores are now included (§7), but reference-based evaluation remains impossible on this test set: the organizers never released reference translations to participants, and the official metric is itself a reference-free LLM assessment. We therefore report no BLEU (Papineni et al., 2002), chrF (Popović, 2015) or COMET (Rei et al., 2020) score, and no correlation between our internal judge and a human gold standard, because neither we nor any other participant can compute them. Both evaluations in this paper are LLM-based, from different model families and rubrics; agreement between them is not evidence of agreement with human translators.

No comparison to quality-aware decoding.

We do not compare our prompted selector against minimum Bayes risk decoding (Freitag et al., 2022) or learned quality-aware reranking (Fernandes et al., 2022) over the same candidate pool (§2). This is the most substantial missing baseline in the paper.

Critic recall is unmeasured. We measure critic precision, on synthetic errors (§5) and on the cues it edits (§6.5), but we never measured recall against naturally occurring errors, which the official defect records (§7) suggest is the more informative quantity for this system. The 233 severity-5 defects in those records form a labeled set on which that recall could be measured directly, and it is the evaluation we would run first on any revised version of the critic.

References

- Rajen Chatterjee, Markus Freitag, Matteo Negri, and Marco Turchi. 2020. Findings of the WMT 2020 shared task on automatic post-editing. In *Proceedings of the Fifth Conference on Machine Translation (WMT)*.
- Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André F. T. Martins, Graham Neubig, Ankush Garg, Jonathan H. Clark, Markus Freitag, and Orhan Firat. 2023. The devil is in the errors: Leveraging large language models for fine-grained machine translation evaluation. In *Proceedings of the Eighth Conference on Machine Translation (WMT)*.
- Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and André F. T. Martins. 2022. Quality-aware decoding for neural machine translation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412.
- Markus Freitag, George Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. 2021. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Transactions of the Association for Computational Linguistics*, 9:1460–1474.
- Markus Freitag, David Grangier, Qijun Tan, and Bowen Liang. 2022. High quality rather than high model probability: Minimum Bayes risk decoding with neural metrics. *Transactions of the Association for Computational Linguistics*, 10:811–825.
- Google. 2026. gemma-4-12B-it-qat-w4a16-ct. <https://huggingface.co/google/gemma-4-12B-it-qat-w4a16-ct>. INT4 quantization-aware-trained release, Apache-2.0. Accessed 2026-08-04.
- Christian Hardmeier. 2015. Pronoun translation and prediction: From classical machine translation to neural networks. In *Proceedings of the Second Workshop on Discourse in Machine Translation*.
- Alina Karakanta, Matteo Negri, and Marco Turchi. 2020. Is 42 the answer to everything in subtitling-oriented speech translation? In *Proceedings of the 17th International Conference on Spoken Language Translation (IWSLT)*.
- Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. *Proceedings of the 24th Annual Conference of the European Association for Machine Translation (EAMT)*.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP)*.
- Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegrefe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-Refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- 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 (ACL)*, pages 311–318.

- Maja Popović. 2015. chrF: Character n -gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation (WMT)*, pages 392–395.
- Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702.
- Hongbin Ren, Lulu Deng, Shaolin Zhu, and Deyi Xiong. 2026. AdaDPI: Document-level translation adaptive agent via dynamic parametric internalization. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 37912–37929.
- Tencent Hunyuan. 2026. Hy-MT2-7B. <https://huggingface.co/tencent/Hy-MT2-7B>. Open-weight translation model, Apache-2.0. Prompt-template slots for background-informed and style-controlled translation are documented under “Hy-MT2 translation task instruction examples”. Accessed 2026-08-04.
- Jörg Tiedemann and Yves Scherrer. 2017. Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*.
- vLLM Project. 2026. vLLM v0.23.0. <https://github.com/vllm-project/vllm/releases/tag/v0.23.0>. Software release; the version served in this work. Accessed 2026-08-07.
- Elena Voita, Rico Sennrich, and Ivan Titov. 2019. When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Ke Wang, Xiutian Zhao, Yanghui Li, and Wei Peng. 2023. PROSE: A pronoun omission solution for Chinese–English spoken language translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2297–2311.
- Yutong Wang, Jiali Zeng, Xuebo Liu, Derek F. Wong, Fandong Meng, Jie Zhou, and Min Zhang. 2025. Delta: An online document-level translation agent based on multi-level memory. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR)*, pages 15708–15731.
- Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2024. Self-preference bias in LLM-as-a-judge. In *NeurIPS 2024 Workshop on Safe Generative AI*. ArXiv:2410.21819.
- WMT26 Organizers. 2026. WMT2026 video subtitle translation shared task. <https://www2.statmt.org/wmt26/video-subtitle-translation.html>. Accessed 2026-08-04.
- WMT26 Video Subtitle Translation Organizers. 2026. Preliminary findings of the WMT26 video subtitle translation shared task: Context and candidate selection matter. Shared task findings paper. Preliminary release, August 2026; official evaluation scores and per-system analysis.
- Minghao Wu, Jiahao Xu, Yulin Yuan, Gholamreza Haffari, Longyue Wang, Weihua Luo, and Kaifu Zhang. 2025. (Perhaps) beyond human translation: Harnessing multi-agent collaboration for translating ultra-long literary texts. *Transactions of the Association for Computational Linguistics*, 13:901–922.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- Mao Zheng, Zheng Li, Tao Chen, Bo Lv, Mingrui Sun, Mingyang Song, Jinlong Song, Hong Huang, Decheng Wu, Hai Wang, Yifan Song, Yanfeng Chen, and Guanwei Zhang. 2026. Hy-MT2: A family of fast, efficient and powerful multilingual translation models in the wild. <https://arxiv.org/abs/2605.22064>. Preprint, arXiv:2605.22064.

A Prompts

The main production prompt templates are reproduced below, in call order. Runtime substitutions are shown as *italic slots*. *bg* is the per-file background block: the series synopsis, the episode identifier, and any character-name spellings already established in the file. Cue-specific context is separate, and appears as *prev* and *next* below. Blank lines present in the originals are elided for space; nothing else is altered. The 12B calls request strict JavaScript Object Notation (JSON); the Hy-MT2 translation calls return plain English text. Decoding: Stage 1 draw 1 at $T=0.7$, draws 2–3 at $T=0.95$ and $T=1.0$ with $\text{top-}p=0.95$, $\text{top-}k=40$, 256 max tokens; all 12B calls at $T=0$, reasoning mode *off* for select and the audit calls and *on* for re-verify and repair, 512–1024 max tokens.

A.1 Stage 1: translate (7B)

In Hy-MT2’s own documented template, which is written in Chinese. Only the target cue is inside 【待翻译文本】 (“text to be translated”); the neighbors live in 【背景信息】 (“background information”) as prose, which is the boundary-preserving structure of §4.3. Line breaks below are inserted to fit the column.

【背景信息】*bg*这句字幕是连续对白的一个片段，上文为：[*prev*]，

下文为: [*<next>*]]. 只翻译【待翻译文本】中的这一句, 不要翻译上下文, 也不要把上下文的内容并入译文。

请结合背景信息将【待翻译文本】翻译为 English, 翻译的风格要严格符合【简洁自然、易读的影视字幕】。注意**只需要输出翻译后的结果, 不要额外解释**:

【待翻译文本】*<source cue>*

In English the prompt reads: “[Background information] *<bg>* This subtitle line is one fragment of continuous dialogue; the preceding line is [*<prev>*] and the following line is [*<next>*]]. Translate only the single line in [text to be translated]; do not translate the surrounding context, and do not merge the content of the context into the translation. Using the background information, translate [text to be translated] into English, in a style that strictly matches [concise, natural, readable screen subtitles]. Note: output only the translated result, with no additional explanation. [text to be translated] *<source cue>*”

The style slot is genre-*neutral* (“concise, natural, readable screen subtitles”); register is carried by the synopsis rather than a hardcoded genre claim. The *retranslate* variant is this same prompt with 注意: (“note:”) followed by *<hint>*, inserted immediately before 【待翻译文本】, where *<hint>* is the concatenation of the per-rule hints of §A.6 for the confirmed defects, most severe first.

A.2 Stage 2: select and minimally edit (12B)

You are the senior editor of English subtitles for a Chinese historical costume drama. A translation model produced several candidate renderings of ONE subtitle line. Choose the best rendering, and lightly edit it for: consistency with established names, correct pronoun gender, period-appropriate register, and BREVITY (must fit on screen). Do NOT add meaning not present in the candidates, and do NOT pull in content from the surrounding context lines --- they are only for your understanding.

Background: *<bg>*

Established character-name spellings (keep consistent): *<names>*

Preceding context (Chinese, do not translate): [*<prev>*]

Following context (Chinese, do not translate): [*<next>*]

TARGET line (Chinese): *<source cue>*

CANDIDATE English renderings:

1. *<candidate 1>* ...

Return ONLY JSON, reasoning first:

```
{"reasoning": "<brief>", "line": "<the chosen/edited English line>", "new_names": ["<any English character names you committed to>"]}
```

The register cue in the opening line is a fixed string

applied to every file, including the modern and variety ones.

A.3 Stage 3: audit call wrapper (12B)

One call per rule; *<defect question>* is substituted from Table 7.

Background: *<bg>*

You are auditing ONE English subtitle line against its Chinese source, for a single specific defect. Answer only about THIS defect.

DEFECT TO CHECK: *<defect question>*

Preceding context (Chinese, for reference only): [*<prev>*]

Following context (Chinese, for reference only): [*<next>*]

SOURCE (Chinese): *<source cue>*

ENGLISH draft: *<en>*

Reason briefly, then answer. A CLEAN draft is the common case --- only report a defect if it is genuinely present. Return ONLY JSON, reasoning first:

```
{"reasoning": "<brief; if defect, quote the offending words>", "defect": true|false, "inverts_meaning": true|false}
```

The *inverts_meaning* field escalates meaning_preserved from major to critical. A ninth rule, the 84-character line ceiling, runs in code and is trusted without re-verification.

A.4 Stage 3: skeptical re-verify (12B, system prompt)

This is the verification pass whose effect §5 measures. It is deliberately asymmetric.

You are an independent senior reviewer of English subtitles. A prior reviewer FLAGGED a line; treat that as an accusation, not a fact. Re-check the rule from scratch. DEFAULT to 'no defect'. Confirm a defect ONLY if you can point to CONCRETE OFFENDING TEXT in the draft that actually violates the rule. If you cannot cite offending text, or the draft is a legitimate rendering, the flag was a false positive --- clear it. Return STRICT JSON only, reasoning first: {"reasoning": "<one sentence; if confirming, quote the offending text>", "defect_confirmed": true|false}.

The user turn supplies RULE FLAGGED, the prior reviewer's reason, the neighbor context, the source and the draft, and closes with “Independently decide whether a REAL defect exists.” The same prompt re-verifies the *repaired* line after repair. On re-verify call failure the original flag is retained, favoring error detection over avoidance of unnecessary repair.

A.5 Stage 4: consolidated repair (12B)

You are the senior editor of English subtitles for a Chinese historical costume drama. The current line has the defects listed below, ranked most to least severe. Produce

Rule key	<i><defect question></i> , verbatim
meaning_preserved	Does the English MISREPRESENT the source line — a hallucinated detail not in the source, a dropped essential meaning, a negation flip, or a who-did-what-to-whom error? (Escalate to a meaning INVERSION if the core claim is reversed.)
pronoun_gender_resolved	Chinese drops subjects and 他/她/它 ('he'/'she'/'it', all pronounced <i>tā</i>) sound alike. Given the context, is the English subject/pronoun the WRONG gender or person for the referent it must denote? (A subject/pronoun that English grammar requires but Chinese omitted is NOT a defect if it matches the context.)
no_context_leak	Does the English contain content that belongs to the NEIGHBOR cues (shown as context) and is NOT in the target line itself?
honorific_register_consistent	Is an honorific/title rendered WRONGLY for the period register (e.g. royal 'We' flattened to 'I', or a rank title dropped/modernized)?
sentence_mood_match	Does the English change the SENTENCE MOOD vs the source — turning a question into a statement or a statement into a question, or flipping a command into a description? (The source's illocutionary force must survive.)
direction_agent_correct	Is the DIRECTION or AGENT reversed — moved-here rendered as moved-there, give/take swapped, or the doer and receiver of the action exchanged?
person_attribution_correct	Is the wrong PERSON doing/saying it — a 3rd-person report ('he will go...') rendered as 1st-person ('I will go...'), or a vocative/address ('Bro, ...') rendered as a described subject ('his brother ...')?
no_invented_name	Does the English state a specific PERSON'S NAME (a romanized given/family name) that is NOT present in the source cue — where the source used only a title, role, or pronoun (e.g. 顾总/王爷/他/她/老板: 'Director Gu', 'my lord', 'he', 'she', 'boss')? Rendering a title AS a title is correct; INVENTING or substituting a proper name the source did not give is the defect. (A name that IS in the source is fine.)

Table 7: The eight audit questions, verbatim. Each is asked in its own call inside the wrapper of Appendix A.

ONE corrected line that resolves ALL of them with the MINIMAL English-side edit. Prefer picking/assembling from the candidate renderings; do NOT translate afresh from the Chinese and do NOT add meaning not supported by the source or candidates. Keep it concise (fit on screen).
Background: *<bg>*
Established names (keep consistent): *<names>*
Preceding context (Chinese): *<[prev]>*
Following context (Chinese): *<[next]>*
SOURCE (Chinese): *<[source cue]>*
CURRENT line: *<en>*
DEFECTS (fix all, worst first):
- *<[severity]>* *<rule>*: *<reasoning>* ...
CANDIDATE renderings to draw from:
1. *<candidate 1>* ...
Return ONLY JSON, reasoning first:
{"reasoning": "<brieft>", "line": "<corrected English line>"}

The candidate pool at this point is $\{\text{initial candidates}\} \cup \{\text{hint-guided retranslations}\}$. If fewer than two distinct strings exist, the call is not made at all: the defects are recorded as residuals and the line is left untouched.

A.6 Rule-to-hint templates (7B)

The hint sent back to the translator, keyed by confirmed rule. These are what make retranslation *informed* rather than blind resampling.

meaning_preserved: Translate only what the source says --- do not add, drop, or reverse meaning.
pronoun_gender_resolved: Use the pronoun/gender that matches the referent in context.
no_context_leak: Translate ONLY this line;

do not pull in content from the surrounding lines.
honorific_register_consistent: Render any title or form of address at a register that matches the source, and keep it consistent --- do not flatten, drop, or modernize it.
sentence_mood_match: Preserve the sentence type of the source: if it is a question, render it as an English question; if a command, a command.
direction_agent_correct: Keep the direction and the doer/receiver of the action exactly as in the source --- do not reverse who acts on whom or which way something moves.
person_attribution_correct: Use the correct grammatical person: if the line REPORTS what another person said or will do, use third person (he/she/they), not 'I'; keep a vocative (a term of address) as address, not as a described subject.
no_invented_name: Do NOT introduce a specific person's name that the source did not give: if the source uses only a title/role/pronoun, render that title/role/pronoun, not an invented proper name.

B Configuration details

Models. tencent/Hy-MT2-7B and google/gemma-4-12B-it-qat-w4a16-ct, both Apache-2.0, used at their released revisions. No adapters, no tokenizer or chat-template changes, no new weights released.

Serving. SageMaker-native vLLM deep learning container (DLC) vllm:0.23.0 (vLLM Project, 2026), one model per endpoint, one 24 GB GPU per endpoint (ml.g5.xlarge /

ml.g6.xlarge; A10G / L4). Client fan-out is 16 cue workers per file with 4 nested workers for the per-rule audit calls, and a per-endpoint in-flight cap that scales with the replica count. The cap must scale with replicas, or the client-side semaphore throttles the whole endpoint to one replica’s throughput. Retries are exponential (2/4/8/16 s) on *transient* failures only (model errors, service-unavailable, throttling, timeouts, 5xx, connection resets), while a clean empty completion is deliberately *not* retried, since that is a model-output issue rather than an infrastructure one.

Inference volume. Over the production run the 7B was called 129,837 times: 125,511 translation draws, three per cue, plus roughly 4,300 hint-guided retranslations. The 12B was called $\approx 366,000$ times: 23,442 candidate selections, 334,696 single-rule audit calls (eight for each of the 41,837 cues), and roughly 8,000 re-verification and repair calls on the cues that were flagged.

Judge protocol. gpt-5.6-luna at high reasoning effort, one item per call, temperature default; the blinding and counterbalancing are described in §6. Line wrapping is stripped from the judged string, since a line break would identify the arm. Judgment loss to malformed JSON was 2 of 3,258 items (0.06%), in one non-headline comparison.

Labeled critic set. The 600 cases of §5 are built from a fixed seed over the 84 non-development files, so the whole set is re-runnable. The 200 controls are half sampled at random from shipped output and half drawn from a cleaner pool. Two rules cannot fill their genre quota, which is a property of the corpus rather than of the sampler: `honorific_register_consistent` and `no_invented_name` have almost no eligible cues in the reality/variety and modern-urban classes.