BREAKING BABEL: A SELF-EVOLVING MULTI-AGENT SYSTEM FOR LONG-FORM SUBTITLE TRANSLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Long-form subtitle translation presents challenges beyond those of conventional machine translation: a single episode may contain hundreds of sentences whose meanings depend on long-range discourse and cultural context spanning the entire episode or even series, while translation quality also requires maintaining consistent terminology and style throughout the series. Existing approaches remain limited. Single-LLM methods operate at the sentence level, lacking long-context understanding and consistent terminology across episodes. Multi-agent methods often rely on static workflows that fail to adapt to scene complexity. In addition, both paradigms ignore the production context, e.g., genre. We propose **SMART**, a **S**elf-evolving **M**ulti-**A**gent system for long-fo**R**m subtitle **T**ranslation. SMART operates in two stages: during test-time training, it builds a persistent series-level memory and iteratively translates a subset of sentences through a dynamic graph router and a Mixture-of-Agents translation layer, equipped with tool-calling modules for terminology verification, subtitle constraint validation, and contextual retrieval; a judge-refiner loop scores each candidate translation and back-propagates textual critiques that refine agent prompts and the routing policy, without retraining the underlying LLMs. During test-time inference, this evolved configuration translates the remaining sentences of the series. To evaluate long-form subtitle translation at scale, we introduce **Subtitle Arena**, covering 14 genres, 2–198 episodes per series, production years from 1959 to 2023, and 15 target locales, together with **SubMQM**, a subtitle-adapted MQM framework with seven dimensions and 19 error categories. SMART achieves the best overall MQM score in all 15 Subtitle Arena directions, reducing the average penalty by 6.9% over the strongest competing agent system. On a public benchmark, MuSC, SMART obtains the best model result across all 4 language pairs. SMART also achieves the best result in human evaluation with an overall score of 4.50/5.

Epigraph: Babel scattered one tongue into many; this work gathers them back.

1 INTRODUCTION

Large language models (LLMs) have substantially advanced machine translation (Singh et al., 2025; Yang et al., 2025; Anthropic, 2026b) in general-domain settings (Xu et al., 2024; 2025a). Long-form subtitle translation, however, remains more challenging than translating isolated sentences. Subtitles often contain hundreds of sentences (Karakanta et al., 2020), while their interpretation often depends on discourse spanning speakers, scenes, and episodes, together with visual and cultural context (Wu et al., 2025). At the same time, translations must preserve terminology and narrative coherence while satisfying strict presentation constraints such as line length and reading speed (Papi et al., 2023; Karakanta et al., 2020).

Existing approaches mainly follow two paradigms. One line adapts a single LLM through subtitle-specific fine-tuning (Cui et al., 2026b) or context-aware prompting (Pramodya et al., 2025) by incorporating neighboring dialogue, genre, or plot summaries. Another line adopts LLM-based multi-agent systems, where specialized roles collaboratively generate and refine translations (Wu et al., 2024), sometimes augmented with multimodal context (Lu et al., 2025). Despite their effectiveness, these approaches remain limited when translation extends over long-form narrative content.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

In particular, we identify three challenges. **First, long-range understanding and consistency**. Single LLM methods (Anthropic, 2026a) primarily reason over local sentences and lack persistent series-level context, causing recurring entities, terminology, and referring expressions to drift across scenes and episodes. **Second, static coordination workflow.** Multi-agent systems often fix the roles and refinement procedures in advance, although translation difficulty varies considerably across scenes. Simple dialogue may require little coordination, while ambiguous references, culturally grounded expressions, or domain-specific scenes may benefit from additional specialists and verification. **Third, contextual adaptation.** Appropriate wording depends strongly on production context, including genre, historical setting, and cultural background, yet current systems have limited ability to retrieve and incorporate such signals dynamically.

To address these challenges, we introduce **SMART**, a **S**elf-evolving **M**ulti-**A**gent system for long-fo**R**m subtitle **T**ranslation. SMART contains two stages: *test-time training* and *test-time inference*. During *test-time training*, SMART translates a subset of a series while evolving its agent prompts, routing policy, and persistent series-level memory. A dynamic graph router selects specialized translators and tools for each sentence, while a Mixture-of-Agents layer generates complementary translation candidates. Tool-calling modules provide terminology verification, subtitle-constraint checking, and contextual retrieval for signals such as genre and production background. A judge-refiner loop evaluates candidate translations and converts observed errors into textual feedback that updates the prompts and routing policy without modifying the underlying LLM parameters. During *test-time inference*, the evolved configuration is frozen and applied to the remaining content, while the series memory continues to accumulate.

To evaluate SMART, we construct **Subtitle Arena** to address limitations of existing subtitle benchmarks, including noisy or misaligned sentences (Cui et al., 2026b; Tiedemann & Luo, 2026) and limited high-quality multilingual pairing (AI, 2024). Beyond sentence-level translation quality, Subtitle Arena is designed to evaluate three properties central to long-form subtitle translation: (i) discourse-aware translation that uses long-range context to resolve the current sentence, (ii) consistent terminology, character references, and style across extended narrative horizons, and (iii) compliance with subtitle-specific display constraints. It contains TV series spanning 14 genres, production years from 1959 to 2023, and 2–198 episodes per series across 15 target locales. We further introduce **SubMQM**, a subtitle-adapted automatic MQM (Multidimensional Quality Metrics) framework covering seven dimensions and 19 fine-grained error categories. Our main contributions are as follows:

- We propose **SMART**, a self-evolving multi-agent framework for long-form subtitle translation with a test-time training stage that jointly adapts agent prompts and routing policies, while maintaining persistent series-level memory and dynamically invoking contextual and verification tools, and a test-time inference stage that translates the rest of the content.
- We introduce **Subtitle Arena**, a long-form subtitle translation benchmark spanning diverse genres, production eras, episodes, and language pairs, with explicit support for evaluating cross-episode consistency and contextual adaptation. To complement subtitle translation metrics, we further adopt a subtitle-adapted MQM framework, **SubMQM**, for fine-grained analysis of semantic, linguistic, contextual, and subtitle-specific technical errors.
- Extensive evaluations show that SMART achieves the best Overall MQM score in all 15 Subtitle Arena directions with a 6.9% lower average penalty than the strongest agent baseline, and the best model result across all 4 MuSC language pairs. SMART also achieves the best result in human evaluation at 4.50/5.

2 Related Work

Multi-Agent Systems and Translation Agents. Multi-agent LLM systems decompose complex tasks across specialized agents (Li et al., 2023; Hu et al., 2021; Guo et al., 2024; Cheng et al., 2024), and have increasingly been adopted for machine translation. Existing systems mainly use role-based collaboration for generation, evaluation, and refinement (He et al., 2024; Wu et al., 2024; Feng et al., 2025; Wang et al., 2025b; Li et al., 2025; Wang et al., 2025a), or introduce persistent context for document and subtitle translation (Wang et al., 2025d; Lu et al., 2025). ALPO (Cui et al., 2026b) instead improves expressive subtitle translation through preference optimization. Despite these advances, most methods operate with predefined workflows or fixed agent behaviors.

Self-Evolving Agents and Prompt Optimization. Self-evolving agents improve their workflows (Hu et al., 2025; Zhang et al., 2025b; Wang et al., 2025c; Zhang et al., 2025a), reusable skills and experience (Wang et al., 2023; Zhao et al., 2024), or persistent memory (Suzgun et al., 2025). Related prompt-optimization methods iteratively improve instructions using search, reflection, or textual feedback (Zhou et al., 2022; Pryzant et al., 2023; Yuksekgonul et al., 2024; Agrawal et al., 2025). However, these approaches typically optimize performance on future tasks or fixed objectives, rather than adapting during a single long-form task while preserving previously acquired behavior.

LLM-as-a-Judge. LLMs are increasingly used to evaluate model outputs through scoring, ranking, and natural-language critiques (Zheng et al., 2023). For machine translation, LLM-based evaluators such as GEMBA-MQM (Kocmi & Federmann, 2023) and AutoMQM (Fernandes et al., 2023) provide structured feedback over translation errors and their severity. Such methods primarily use the judge for evaluation or local refinement.

Key Differences. SMART differs from prior work in three respects. Unlike translation agents with fixed workflows, SMART evolves at test time, adapting agent instructions, routing, and contextual knowledge. Unlike general self-evolving agents, SMART targets the long-form translation task, improving future sentences while preserving terminology, style, and prior translation competence. Unlike conventional prompt optimization, SMART treats prompts and agent behaviors as persistent system state updated through translation-specific feedback.

3 METHODOLOGY

3.1 PROBLEM FORMULATION

Let an episode-level subtitle be $\mathcal{D} = \{d_i\}_{i=1}^{N}$, where each sentence $d_i = (x_i, \tau_i, b_i)$ contains source text x_i, a timestamp interval τ_i, and display constraints b_i such as characters per line and reading speed. Given a target locale L, the goal is to generate $\mathcal{Y} = \{y_i\}_{i=1}^{N}$ that is locally faithful while remaining globally coherent across episodes of the same series. The translation must preserve meaning, character narrative voice, terminology, and subtitle-specific display constraints over long contexts.

At training step t, SMART maintains the system state

$$\mathcal{S}_t = (\mathcal{R}, \Pi_t, \rho_t, M_t), \tag{1}$$

where $\mathcal{R}$ is a fixed pool of specialized translator roles, $\Pi_t = \{\pi_{r,t}\}_{r\in\mathcal{R}}$ denotes their prompts, ρ_t is a routing policy that selects active translators, and M_t is a persistent series-level memory. Unlike fixed multi-agent pipelines, SMART adapts both how translators are instructed and which translators are invoked from translation feedback, without updating the underlying LLM parameters.

3.2 SYSTEM OVERVIEW

Fig. 1 summarizes SMART through three interacting mechanisms. First, a **persistent memory pool** maintains series-level information that should survive beyond individual model calls, including terminology, character profiles, domain knowledge, and target-language idioms. Second, an **adaptive translation graph** routes each subtitle sentence to a subset of specialized translators, which may invoke tools for contextual retrieval, terminology control, external knowledge, and subtitle validation. Third, a **judge–refiner loop** evaluates the resulting candidates, selects the most promising translation, and refines it before the accepted result is written back to memory.

For **test-time training and inference**, we split the episodes of each series into $\mathcal{D}_{\text{tr}}$ and a held-out set $\mathcal{D}_{\text{inf}}$ using a 3:7 ratio. Translation feedback from $\mathcal{D}_{\text{tr}}$ is used during training to improve translator prompts and routing decisions. These components are then frozen for $\mathcal{D}_{\text{inf}}$ during inference, while memory continues to grow as translation proceeds across episodes. Appendix B.7 details the optimization protocol, and Appendix C.1 provides an end-to-end example.

3.3 PERSISTENT SERIES MEMORY

Long-form subtitle translation contains information that is sparse but persistent: a character, expression, or domain-specific term may be introduced in one episode and recur much later. SMART

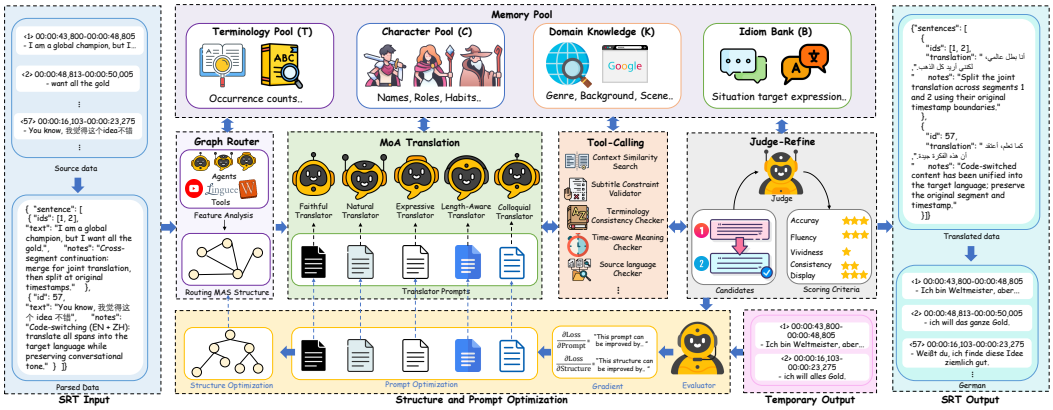


Figure 1: Overview of **SMART**. Given source data in SRT style, SMART parses it into sentences. For each sentence, a learnable router (**Graph Router**) activates a subset of translators (**MoA Translation**); candidates may invoke tools (**Tool-Calling**) and are selected and refined by a judge–refiner loop (**Judge-Refine**). A persistent **Memory Pool** carries terminology, character information, domain knowledge, and idiomatic expressions across episodes. During test-time training, translation feedback updates translator prompts and routing.

therefore maintains a series-level memory

$$M = (T, C, K, B), \tag{2}$$

where T is a terminology pool, C stores character profiles, K stores domain and scene knowledge, and B is an idiom bank. Before translating episode e, SMART loads the state $M^{(e)}$ accumulated from earlier episodes. Newly confirmed information is merged back after translation:

$$M^{(e+1)} = M^{(e)} \cup \Delta M^{(e)}. \tag{3}$$

All translators and tools can query this memory, allowing recurring entities and prior translation decisions to be retrieved rather than independently re-derived. The memory therefore provides an explicit long-range state beyond the context window of any individual model call. Appendix C.3 provides a concrete cross-episode example.

SMART additionally constructs a lightweight content profile containing the series genre, setting, principal characters, and domain background. Ambiguous jargon, cultural references, and slang may trigger external retrieval. The retrieved evidence is interpreted in the context of the current scene before being stored, separating literal external knowledge from the register and wording appropriate for subtitle translation. Further details are provided in Appendix B.

3.4 Adaptive Multi-Agent Translation

Dynamic routing. Rather than invoking every translator for every sentence, SMART selects an active set $\mathcal{A}_i \subseteq \mathcal{R}$ according to

$$\mathcal{A}_i = \rho_t(\phi(d_i, M_t)), \tag{4}$$

where $\phi(\cdot)$ summarizes properties relevant to translation, including sentence length, scene tone, and lexical cues such as slang or profanity. The translator pool contains complementary roles emphasizing semantic faithfulness, naturalness, expressiveness, subtitle-length control, and colloquial rendering. The router can therefore allocate additional translation capacity when a sentence benefits from it instead of applying a fixed topology to every input.

Mixture-of-agents translation with tools. Each active translator $r \in \mathcal{A}_i$ produces a candidate under its current prompt $\pi_{r,t}$:

$$y_i^{(r)} = r(d_i, \pi_{r,t}, M_t, \mathcal{T}), \tag{5}$$

where $\mathcal{T}$ denotes the tool set. These tools expose information that is better retrieved or verified on demand than embedded in a monolithic prompt, including surrounding context, confirmed terminology, similar translated sentences, subtitle constraints, domain knowledge, idiom lookup, and target-language fluency checks. Appendix B.4 documents the tool interfaces and execution modes.

Judge and refinement. The judge evaluates each candidate along semantic accuracy, fluency, style and register, long-range consistency, and subtitle display constraints:

$$c_i^{(r)}, \delta_i^{(r)} = J\Big(y_i^{(r)}, d_i, M_t\Big), \tag{6}$$

where $c_i^{(r)}$ is the score and $\delta_i^{(r)}$ the corresponding critique. SMART selects the highest-scoring candidate and refines it using this feedback:

$$r^\star = \arg\max_{r \in \mathcal{A}_i} c_i^{(r)}, \qquad \hat{y}_i = R\Big(y_i^{(r^\star)}, \delta_i^{(r^\star)}, M_t\Big). \tag{7}$$

The accepted translation updates the episode history and persistent memory. After each episode, an episode-level pass revisits the translated subtitles to repair residual consistency errors that are difficult to detect from a single sentence alone. Appendix C.1 illustrates the complete sentence-level decision path.

3.5 Self-Evolution at Test-Time Training

SMART evolves its translation process from feedback collected on $\mathcal{D}_{\mathrm{tr}}$. After each training batch $\mathcal{B}_t$, an evaluator aggregates candidate scores and textual critiques into a structured feedback signal:

$$\Delta_t = E\left(\left\{c_i^{(r)}, \delta_i^{(r)}\right\}_{i \in \mathcal{B}_t,\, r \in \mathcal{A}_i}\right). \tag{8}$$

This feedback updates both translator prompts and the routing policy:

$$\Pi_{t+1} = F^{\Pi}(\Pi_t; \Delta_t), \qquad \rho_{t+1} = F^{\rho}(\rho_t; \Delta_t). \tag{9}$$

F^{Π} revises translator prompts to address recurring failure modes, while F^{ρ} adjusts the router toward translator configurations that receive higher scores on similar inputs. Both updates are expressed in natural language and occur entirely at test time; the underlying LLM parameters remain fixed. Appendix C.5 and Appendix C.6 provide examples of the two update processes.

Both are constrained rewrites, not free-form edits: F^{Π} revises an underperforming agent's prompt while preserving its role and workflow, and F^{ρ} rewrites a finite table from routing categories to agent and tool subsets under a floor of two agents per category. Both apply once per training epoch from batch-aggregated feedback rather than per sentence. Appendix B.8 gives the trigger conditions and constraints.

After T training steps, SMART freezes (Π_T, ρ_T) and applies the resulting configuration to $\mathcal{D}_{\mathrm{inf}}$. Self-evolution thus changes the translation policy only during test-time training, whereas the memory M continues updating during inference to preserve cross-sentence and cross-episode consistency. This separation allows SMART to adapt its translation strategy without leaking held-out feedback into the frozen inference configuration. Appendix B.7 gives the complete optimization loop.

4 Subtitle Arena: Benchmark and Evaluation

4.1 Benchmark Construction

Benchmark scope. Existing subtitle benchmarks are not designed to fully capture the challenges of long-form viewing, including discourse preservation, terminology and character consistency, and subtitle-specific display constraints. We therefore introduce **Subtitle Arena**, a series-centric benchmark containing 192 television series, 6,267 English source episodes, and 70,664 aligned bilingual episode pairs across 15 target locales.

Table 1: Statistics of **Subtitle Arena**. **Top:** aligned bilingual episode coverage for each target locale, measured against 6,267 English source episodes. **Bottom:** distribution of the 192 television series across 14 genres.

Aligned Episode Coverage by Target Locale								
Locale	zh-CN	pt-BR	es-ES	es-MX	fr-FR	ro-RO	tr-TR	pt-PT
Episodes	6,267	5,407	5,144	5,144	5,079	5,002	4,956	4,853
Locale	de-DE	it-IT	nl-NL	sv-SE	da-DK	no-NO	ko-KR	**Total**
Episodes	4,697	4,518	4,518	4,286	4,273	3,277	3,243	**70,664**

Series Distribution by Genre								
Genre	Comedy	Action	Crime	Animation	Fantasy	Drama	Sci-Fi	
Series	20	20	20	18	15	15	15	
Genre	Documentary	Romance	Mystery	Horror	Adventure	Thriller	Reality	**Total**
Series	12	12	10	10	10	10	5	**192**

As shown in Table 1, Subtitle Arena provides broad coverage across 15 target locales and 14 television genres. Six locales contain more than

5,000 aligned episodes, and all locales retain more than 3,200 episode pairs. Its series-level organization supports evaluation over continuous narrative contexts, while the diversity of languages and genres tests robustness across different discourse, stylistic, and locale-specific conditions. Benchmark statistics are provided in Appendix D.6.

Dataset construction. Subtitle Arena is constructed from OpenSubtitles2024 (Tiedemann & Luo, 2026), which contains subtitles released before 2024. Rather than treating aligned sentences as independent translation examples, we reorganize the data at the episode and series levels. Specifically, we recover bilingual correspondences and associate subtitle files using IMDb identifiers, season indices, and episode indices. For each subtitle cue, we preserve the original start and end timecodes, normalize malformed formatting, and remove empty, one-sided, or unparsable alignments. We then group aligned episodes from the same television series to preserve long-range narrative structure. Full details on data processing, locale mapping, and filtering are provided in Appendix D.

Purpose. Subtitle Arena is designed to test three properties that are underrepresented in conventional MT evaluation: (i) discourse-aware translation, where long-range context is needed to interpret and translate the current sentence; (ii) consistency of terminology, character references, and style across extended narrative horizons; and (iii) compliance with subtitle-specific display requirements (Papi et al., 2023; Wilken et al., 2022). Its series-level organization further allows us to test whether translation behavior remains robust across different genres, scenes, and locale-specific conventions.

4.2 LLM-AS-A-JUDGE EVALUATION WITH SUBMQM

For the LLM-as-a-judge evaluation, we develop a new rubric-based evaluation metric, **SubMQM**, a subtitle-adapted Multidimensional Quality Metrics (MQM) protocol. For each hypothesis, an LLM evaluator assigns error penalties along seven dimensions—Terminology, Accuracy, Fluency, Linguistic Conventions, Technical, Locale Conventions, and Audience Appropriateness—which are further decomposed into 19 error types. Each error type receives a penalty in $\{0, 5, 10\}$ for no, minor, and severe error, respectively; dimension-level scores aggregate the corresponding error types, and Overall places greater weight on semantic fidelity. All SubMQM results are therefore reported as penalties, where lower is better. We use the same evaluator and rubric for all systems. The complete rubric, weighting scheme, and alignment procedure are provided in Appendix E.

Table 2: **Overview of SubMQM.** Each error type is assigned a penalty in $\{0, 5, 10\}$; lower is better. The complete 19-type rubric is provided in Appendix E.

Dimension	Representative Errors	#	Weight
Accuracy	Mistranslation, omission, overtranslation	3	0.30
Terminology	Name and term inconsistency	2	0.20
Fluency	Coherence, naturalness, vividness	3	0.20
Audience Appropriateness	Profanity, formality	2	0.12
Linguistic Conventions	Punctuation, capitalization, grammar, spacing	4	0.08
Technical	Line breaking, CPL, lines per box	3	0.06
Locale Conventions	Localization, language detection	2	0.04
Total	**19 error types**	**19**	**1.00**

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Baselines. We compare against three classes of systems. Online refers to community-authored subtitles distributed with the corresponding episodes in OpenSubtitles. Because these subtitles are collected in the wild and vary in quality, we do not treat them as a controlled human upper bound. Single-call LLMs include Gemma 3 4B, DeepSeek-V3.2, Claude Sonnet 4.6, Claude Opus 4.8, and GPT-5.5. These models receive the same local context and subtitle-formatting instructions as SMART, but translate each sentence once without routing, persistent memory, tools, candidate selection, or refinement. Agentic baselines include TransAgent (Wu et al., 2024) and DRT (Wang et al., 2025b), both reimplemented with Claude Sonnet 4.6 for a controlled comparison with SMART. On the public MuSC benchmark (Cui et al., 2026b), we additionally report the systems evaluated in the original benchmark, including the fine-tuned ALPO variant of Qwen2.5-14B.

Implementation. Unless otherwise specified, all SMART roles use Claude Sonnet 4.6. For each series, the first 30% of the available content is used for test-time training and the remaining 70% for held-out inference. Prompt and routing updates are permitted only during training and are frozen thereafter; persistent series memory continues to accumulate because it represents task state. The

episode-level consistency check operates over overlapping windows of size 5, 10, and 20 with 50% overlap. Section 5.3 tests weaker translation backbones and an alternative evaluator.

5.2 Main Results

Results on Subtitle Arena. Table 3 summarizes representative bidirectional results on Subtitle Arena. Complete fine-grained results for all 30 translation directions and all 19 SubMQM error types are provided in Appendix F.1 and Appendix F.2. Across the 15 English→locale directions, SMART obtains the lowest Overall SubMQM penalty in every direction and reduces the mean Overall penalty from 1.20 (TransAgent) to 1.11, a relative reduction of 6.9%. Despite using Claude Sonnet 4.6 as its default backbone, SMART also improves over stronger single-call translators, including GPT-5.5 (1.44) and Claude Opus 4.8 (1.55) on average.

The improvements are distributed across semantic and subtitle-specific dimensions rather than being driven by one error type. Relative to TransAgent, SMART lowers the mean Accuracy penalty from 1.83 to 1.67 and the mean Technical penalty from 1.83 to 1.72, while also reducing Terminology, Fluency, Linguistic Conventions, Locale Conventions, and Audience Appropriateness. The reverse locale→English evaluation shows the same overall trend: SMART reduces the mean Overall penalty from 0.41 to 0.37, a 10.0% relative reduction, and achieves the lowest Overall score in all 15 directions. These results indicate that the gains are not only specific to generating diverse target languages but also persist when every system generates English.

Results on the MuSC dataset. We further evaluate SMART on MuSC to test whether its gains transfer beyond Subtitle Arena. As shown in Table 4, SMART achieves the best model result on all Accuracy, Naturalness, and Vividness measurements across the four translation directions. Compared to the strongest non-SMART result in each column, SMART improves by an average of approximately 3.0 points, with gains ranging from 1.2 to 7.9 points. The advantage remains clear against both strong reasoning models and the fine-tuned ALPO baseline. The largest gains appear in Vividness, where SMART improves over the strongest prior system by 3.8, 7.9, 2.0, and 6.3 points on en→zh, ko→zh, zh→en, and zh→th, respectively. This result is particularly important for subtitle translation, where high-quality outputs require not only semantic fidelity but also contextual and stylistic adaptation over long-form narrative content.

Table 4: MuSC results across four directions. Higher is better. Best and second-best are highlighted in blue-gray and warm beige.

Model	Training	en→zh			ko→zh			zh→en			zh→th		
		Acc.	Nat.	Viv.	Acc.	Nat.	Viv.	Acc.	Nat.	Viv.	Acc.	Nat.	Viv.
Gold Reference	Human	83.6	82.6	71.5	78.0	77.8	65.8	83.0	80.3	73.3	76.6	75.1	66.3
VideoDubber	–	46.9	51.9	49.7	39.6	45.2	48.2	53.6	54.8	50.1	34.1	34.9	41.5
NLLB-3.3B	–	61.4	54.0	43.7	33.1	26.1	25.4	29.1	21.7	20.8	42.6	33.9	40.5
MADLAD-10B	–	59.7	55.5	46.3	44.9	42.9	46.7	45.1	38.9	37.6	47.9	50.8	51.0
Google Translate	–	84.2	79.7	54.4	54.9	52.8	52.0	79.8	66.3	50.2	55.2	56.2	54.5
GPT-4o	ICL (C)	89.3	82.3	59.8	80.0	79.9	58.1	88.5	83.0	64.6	88.0	84.4	67.9
Qwen-Max	ICL (C)	91.9	84.4	61.3	83.7	82.5	61.8	**90.0**	85.0	66.8	91.3	**85.8**	69.1
DeepSeek-V3.1	ICL (C)	91.2	85.3	63.5	83.1	82.2	57.2	89.5	84.1	63.0	89.9	84.6	67.1
DeepSeek-R1	ICL (R)	90.5	85.7	70.8	79.8	81.6	65.6	88.5	85.6	73.5	87.6	84.0	71.0
GPT-5	ICL (R)	**92.4**	**87.0**	71.1	**84.5**	82.6	65.0	89.1	86.1	75.2	88.7	83.9	73.0
Qwen2.5-14B	SFT	86.4	82.0	59.1	80.9	76.1	53.9	85.2	80.1	54.8	87.3	82.6	66.0
Qwen2.5-14B	ALPO	90.6	84.3	**76.6**	84.3	**83.3**	**70.5**	88.3	**86.8**	**81.7**	**91.9**	84.7	**74.2**
SMART	–	**94.3**	**88.2**	**80.4**	**86.8**	**85.9**	**78.4**	**92.3**	**88.6**	**83.7**	**94.2**	**87.0**	**80.5**

Abbreviations. Acc. = Accuracy; Nat. = Naturalness; Viv. = Vividness. SFT = supervised fine-tuning; ICL = in-context learning; (C) = chat model; (R) = reasoning model.

Multimodal Expansion. We extend SMART with video and audio tools powered by Qwen3-Omni (Xu et al., 2025b). Across six representative directions, SMART + MM reduces the average SubMQM penalty from 0.63 to 0.57 and outperforms the multimodal subtitle systems ViDove (Lu et al., 2025) and Hermes (Cui et al., 2026a). Full results are provided in Appendix G.

Temporal Generalization. On 200 TV series released in 2025–2026, including titles beyond the reported knowledge cutoff of Claude Sonnet 4.6, SMART outperforms TransAgent across all six evaluated directions. This suggests that its gains extend to temporally shifted content rather than being confined to Subtitle Arena. Details are in Appendix F.4.

Qualitative Analysis. Rendered examples further illustrate SMART's advantages on context-sensitive translation errors. Details are in Appendix H.

5.3 Backbone and Evaluator Robustness

Table 5 separates the effect of SMART's scaffold from backbone strength on representative bidirectional pairs. For en→zh, SMART reduces the Overall penalty from 1.46 to 0.67 with Gemma 3 4B and from 0.93 to 0.53 with DeepSeek-V3.2; similar reductions hold in the reverse direction and across the other pairs. These results indicate that SMART's gains are not solely due to using a

Table 3: SubMQM results on representative bidirectional language pairs from Subtitle Arena. Each cell reports both directions in the order shown in the first column (left | right). Dimension scores are averaged over their constituent error types. All values are penalties (lower is better). The best and second-best values are highlighted in blue and orange, respectively.

Language Direction	Method	Term.	Acc.	Flu.	Ling.	Tech.	Locale	Audience	Overall
en→zh | zh→en	Online	0.35 | 0.42	4.87 | 4.23	4.07 | 3.28	1.98 | 1.47	0.20 | 0.29	1.40 | 0.11	**0.05** | 0.22	2.58 | 2.17
	Gemma 3 4B	0.49 | 0.52	2.76 | 3.10	1.97 | 2.16	0.59 | 0.81	0.79 | 1.00	0.15 | 0.16	0.35 | 0.39	1.46 | 1.64
	DeepSeek-V3.2	0.28 | 0.31	1.86 | 2.03	1.25 | 1.43	0.29 | 0.42	0.49 | 0.88	**0.03** | 0.08	0.13 | 0.18	0.93 | 1.07
	Claude 4.6	0.26 | 0.27	1.63 | 1.84	1.16 | 1.28	0.26 | 0.34	0.42 | 0.83	**0.03** | 0.05	0.16 | 0.15	0.84 | 0.96
	Claude 4.8	0.25 | 0.20	1.56 | 1.39	1.09 | 1.00	0.20 | 0.23	0.32 | 0.68	0.04 | 0.03	0.13 | 0.11	0.78 | 0.73
	GPT-5.5	0.20 | 0.18	1.16 | 1.20	1.13 | 0.98	0.19 | 0.19	0.22 | 0.87	**0.03** | 0.03	0.14 | 0.13	0.66 | 0.67
	DRT	0.22 | 0.16	1.11 | 1.01	**1.04** | **0.86**	0.12 | 0.15	0.12 | 0.30	0.04 | **0.03**	**0.11** | 0.11	0.61 | 0.55
	TransAgent	**0.16** | **0.14**	**0.85** | **0.92**	1.06 | **0.79**	**0.08** | **0.13**	**0.05** | **0.25**	0.04 | **0.03**	0.14 | **0.10**	**0.52** | **0.50**
	SMART	**0.15** | **0.13**	**0.80** | **0.80**	**0.95** | **0.79**	**0.07** | **0.10**	**0.00** | **0.08**	0.04 | **0.02**	0.12 | **0.09**	**0.48** | **0.44**
en→de | de→en	Online	0.51 | 0.39	6.69 | 3.69	2.76 | 2.45	0.61 | 1.01	**0.05** | 0.23	0.91 | 0.29	**0.16** | 0.19	2.77 | 1.80
	Gemma 3 4B	0.70 | 0.46	4.18 | 2.86	2.74 | 1.92	0.86 | 0.67	1.77 | 0.90	0.48 | 0.19	0.58 | 0.32	2.21 | 1.49
	DeepSeek-V3.2	0.44 | 0.29	2.90 | 1.88	1.92 | 1.28	0.41 | 0.34	1.38 | 0.82	0.22 | 0.21	0.31 | 0.52	1.50 | 1.02
	Claude 4.6	0.41 | 0.25	2.56 | 1.71	1.81 | 1.16	0.41 | 0.28	**1.18** | 0.76	0.19 | 0.08	0.31 | 0.13	1.36 | 0.88
	Claude 4.8	0.40 | 0.18	2.31 | 1.29	1.54 | 0.92	0.37 | 0.19	1.59 | 0.61	0.15 | 0.06	**0.28** | 0.10	1.25 | 0.67
	GPT-5.5	0.36 | 0.16	2.07 | 1.11	1.71 | 0.89	0.32 | 0.16	1.58 | 0.78	0.11 | 0.05	0.38 | 0.11	1.21 | 0.62
	DRT	**0.24** | 0.15	1.62 | 0.94	1.62 | 0.78	0.28 | 0.13	1.59 | 0.27	0.08 | **0.04**	0.35 | 0.09	1.02 | 0.50
	TransAgent	0.26 | **0.13**	**1.51** | **0.85**	**1.46** | **0.71**	**0.27** | **0.11**	1.31 | **0.21**	**0.05** | **0.04**	0.40 | **0.08**	**0.95** | **0.45**
	SMART	**0.25** | **0.12**	**1.24** | **0.74**	**1.46** | **0.71**	**0.22** | **0.08**	1.55 | **0.07**	**0.03** | **0.04**	0.41 | **0.07**	**0.87** | **0.41**
en→ko | ko→en	Online	0.41 | 0.44	2.91 | 4.44	1.87 | 3.40	0.63 | 1.56	0.27 | 0.34	**0.04** | 0.12	0.70 | 0.24	1.48 | 2.28
	Gemma 3 4B	0.68 | 0.54	3.23 | 3.30	2.43 | 2.29	0.64 | 0.87	0.80 | 1.04	0.22 | 0.17	1.04 | 0.42	1.82 | 1.74
	DeepSeek-V3.2	0.42 | 0.33	2.09 | 2.16	1.63 | 1.51	0.31 | 0.44	0.47 | 0.94	**0.04** | 0.08	**0.66** | 0.19	1.17 | 1.13
	Claude 4.6	0.40 | 0.29	2.03 | 1.93	**1.32** | 1.34	0.29 | 0.36	0.44 | 0.87	**0.04** | 0.07	**0.61** | 0.16	1.07 | 1.01
	Claude 4.8	0.35 | 0.22	1.80 | 1.46	1.56 | 1.05	0.27 | 0.25	0.35 | 0.71	**0.04** | 0.05	0.72 | 0.13	1.05 | 0.77
	GPT-5.5	0.37 | 0.19	1.85 | 1.25	1.47 | 1.03	0.22 | 0.21	0.25 | 0.91	**0.04** | 0.05	0.68 | 0.14	1.04 | 0.71
	DRT	**0.28** | 0.17	1.71 | 1.05	1.50 | **0.90**	0.22 | 0.16	0.20 | 0.31	**0.04** | **0.04**	0.75 | 0.11	0.99 | 0.57
	TransAgent	0.32 | **0.15**	**1.63** | **0.96**	1.57 | **0.82**	**0.19** | **0.14**	**0.13** | **0.26**	0.05 | **0.04**	0.67 | **0.11**	**0.97** | **0.52**
	SMART	**0.29** | **0.14**	**1.55** | **0.84**	**1.43** | **0.82**	**0.17** | **0.11**	**0.10** | **0.10**	**0.04** | **0.04**	0.79 | **0.10**	**0.92** | **0.47**
en→it | it→en	Online	0.54 | 0.34	6.36 | 3.40	3.04 | 2.30	2.21 | 0.92	**0.13** | **0.20**	0.68 | 0.22	**0.07** | 0.16	2.85 | 1.66
	Gemma 3 4B	0.70 | 0.42	4.25 | 2.70	3.25 | 1.80	1.01 | 0.62	1.69 | 0.85	0.44 | 0.15	0.62 | 0.28	2.34 | 1.39
	DeepSeek-V3.2	0.43 | 0.26	2.94 | 1.75	2.46 | 1.20	0.55 | 0.31	**1.29** | 0.77	0.20 | 0.09	0.29 | 0.14	1.62 | 0.91
	Claude 4.6	0.38 | 0.22	2.54 | 1.59	2.42 | 1.07	0.53 | 0.25	1.48 | 0.72	0.20 | 0.07	**0.27** | 0.12	1.49 | 0.82
	Claude 4.8	0.37 | 0.16	2.28 | 1.19	2.43 | 0.85	0.43 | 0.17	1.36 | 0.58	0.19 | 0.05	**0.27** | 0.09	1.39 | 0.62
	GPT-5.5	0.32 | 0.15	2.13 | 1.03	2.43 | 0.83	0.39 | 0.14	1.82 | 0.73	0.15 | 0.05	0.31 | 0.10	1.37 | 0.57
	DRT	**0.27** | 0.13	**1.66** | 0.86	2.23 | 0.73	0.37 | 0.11	1.89 | 0.25	**0.13** | **0.04**	0.34 | 0.08	1.19 | 0.46
	TransAgent	**0.21** | **0.12**	1.73 | **0.78**	**2.03** | **0.66**	**0.29** | **0.10**	2.28 | **0.20**	**0.13** | **0.03**	0.33 | **0.07**	**1.17** | **0.42**
	SMART	**0.21** | **0.11**	**1.49** | **0.69**	**2.15** | **0.65**	**0.28** | **0.08**	2.37 | **0.07**	**0.11** | **0.03**	0.35 | **0.06**	**1.13** | **0.38**

Abbreviations. Term. = Terminology; Acc. = Accuracy; Flu. = Fluency; Ling. = Linguistic Conventions; Tech. = Technical. Within each cell, the left and right values correspond to the left and right translation directions in the first column, respectively. Complete fine-grained results for all 19 SubMQM error types and all 30 translation directions are provided in Appendix F.1 and Appendix F.2: English→locale results are reported in Tables 11 and 12, while locale→English results are reported in Tables 13 and 14.

stronger base model. We further replace the default Claude Sonnet 4.6 judge with GPT-5.5 while keeping translation outputs fixed. Overall penalties change only marginally, e.g., 0.48 versus 0.47 on en→zh and 0.87 versus 0.87 on en→de, suggesting that the evaluation is robust to the choice of judge. Full backbone and judge results across all 15 bidirectional language pairs are in Appendix F.3.

5.4 Ablation Study

Table 6 isolates the contribution of SMART's major components across six representative directions. Series-level memory has the largest effect, increasing the Overall penalty by 0.78 on average when removed, followed by MoA (+0.61), contextual retrieval and the idiom bank (+0.56), and sliding-window consistency (+0.47). These results highlight the importance of maintaining long-range context, retrieving relevant knowledge, and reconciling multiple specialized translation hypotheses.

Table 6: Component ablation of SMART across six directions. Values are Overall SubMQM penalties (lower is better); parentheses show increases over the full system.

Setting	en→zh	en→de	en→ko	en→it	en→es	en→fr
Full system	**0.48**	**0.88**	**0.93**	**1.13**	**1.18**	**1.19**
w/o Dynamic Router	**0.70** (+0.22)	**1.12** (+0.24)	**1.18** (+0.25)	1.47 (+0.34)	1.46 (+0.28)	1.48 (+0.29)
w/o Self-Evolution	0.80 (+0.32)	1.22 (+0.34)	1.28 (+0.35)	1.50 (+0.37)	1.55 (+0.37)	1.57 (+0.38)
w/o Memory	1.43 (+0.95)	1.78 (+0.90)	1.69 (+0.76)	1.67 (+0.54)	1.74 (+0.56)	2.13 (+0.94)
w/o Contextual Retrieval & Idiom Bank	1.24 (+0.76)	1.56 (+0.68)	1.31 (+0.38)	1.49 (+0.36)	1.44 (+0.26)	2.08 (+0.89)
w/o Sliding Window	0.93 (+0.45)	1.41 (+0.53)	1.52 (+0.59)	1.52 (+0.39)	1.58 (+0.40)	1.64 (+0.45)
w/o MoA	1.21 (+0.73)	1.66 (+0.78)	1.66 (+0.73)	1.63 (+0.50)	1.61 (+0.43)	1.70 (+0.51)
w/ Combined Prompt	1.06 (+0.58)	1.53 (+0.65)	1.47 (+0.54)	**1.32** (+0.19)	**1.41** (+0.23)	**1.44** (+0.25)

Dynamic routing and self-evolution provide consistent gains, with penalty increases of 0.27 and 0.36 when removed. Replacing MoA with a single translator using a Combined Prompt increases the

Table 5: Backbone and judge analysis on 4 bidirectional pairs from Subtitle Arena. Each cell gives both directions in the first column's order (left | right). All values are penalties (lower is better). Blue-gray and warm-beige mark the best and second-best values.

Language Direction	Method / Setting	Term.	Acc.	Flu.	Ling.	Tech.	Locale	Audience	Overall
en→zh | zh→en	Gemma 3 4B	0.49 | 0.52	2.76 | 3.10	1.97 | 2.16	0.59 | 0.81	0.79 | 1.00	0.15 | 0.16	0.35 | 0.39	1.46 | 1.64
	DeepSeek-V3.2	0.28 | 0.31	1.86 | 2.03	1.25 | 1.43	0.29 | 0.42	0.49 | 0.88	**0.03** | 0.08	**0.13** | 0.18	0.93 | 1.07
	SMART (Gemma 3 4B)	0.23 | 0.19	1.14 | 1.12	1.27 | 1.06	0.11 | 0.16	0.03 | 0.11	0.07 | 0.05	0.19 | 0.12	0.67 | 0.62
	SMART (DeepSeek-V3.2)	**0.18** | 0.15	**0.89** | 0.92	1.04 | 0.91	**0.08** | **0.12**	**0.01** | **0.09**	0.05 | **0.03**	0.14 | **0.10**	0.53 | **0.51**
	SMART (Judge: GPT-5.5)	**0.15** | **0.12**	**0.80** | **0.76**	**0.90** | **0.82**	**0.07** | **0.10**	**0.00** | **0.08**	**0.04** | **0.02**	**0.12** | **0.09**	**0.47** | **0.44**
	SMART (Claude 4.6)	**0.15** | **0.13**	**0.80** | **0.80**	**0.95** | **0.79**	**0.07** | **0.10**	**0.00** | **0.08**	**0.04** | **0.02**	**0.12** | **0.09**	**0.48** | **0.44**
en→de | de→en	Gemma 3 4B	0.70 | 0.46	4.18 | 2.86	2.74 | 1.92	0.86 | 0.67	1.77 | 0.90	0.48 | 0.19	0.58 | 0.32	2.21 | 1.49
	DeepSeek-V3.2	0.44 | 0.29	2.90 | 1.88	1.92 | 1.28	0.41 | 0.34	**1.38** | 0.82	0.22 | 0.21	**0.31** | 0.52	1.50 | 1.02
	SMART (Gemma 3 4B)	0.33 | 0.16	1.68 | 1.10	2.18 | 1.03	0.33 | 0.13	2.15 | 0.10	0.06 | 0.07	0.60 | 0.10	1.23 | 0.60
	SMART (DeepSeek-V3.2)	**0.27** | 0.14	1.41 | 0.88	1.68 | 0.80	**0.26** | 0.10	1.78 | **0.08**	**0.04** | **0.05**	0.45 | **0.08**	**0.99** | **0.47**
	SMART (Judge: GPT-5.5)	**0.25** | **0.11**	**1.22** | **0.78**	**1.51** | **0.68**	**0.22** | **0.09**	**1.51** | **0.07**	**0.03** | **0.04**	**0.39** | **0.08**	**0.87** | **0.41**
	SMART (Claude 4.6)	**0.25** | **0.12**	**1.24** | **0.74**	**1.46** | **0.71**	**0.22** | **0.08**	1.55 | **0.07**	**0.03** | **0.04**	0.41 | **0.07**	**0.87** | **0.41**
en→ko | ko→en	Gemma 3 4B	0.68 | 0.55	3.23 | 3.30	2.43 | 2.29	0.64 | 0.87	0.80 | 1.04	0.22 | 0.17	1.04 | 0.42	1.82 | 1.74
	DeepSeek-V3.2	0.42 | 0.33	2.09 | 2.16	**1.63** | 1.51	0.31 | 0.44	0.47 | 0.94	**0.04** | 0.08	**0.66** | 0.19	1.17 | 1.13
	SMART (Gemma 3 4B)	0.42 | 0.22	2.03 | 1.16	1.96 | 1.23	0.25 | 0.17	0.14 | 0.15	0.07 | 0.07	1.29 | 0.15	1.26 | 0.68
	SMART (DeepSeek-V3.2)	0.33 | **0.17**	1.68 | **0.98**	1.65 | 0.98	0.19 | **0.13**	**0.12** | **0.11**	**0.05** | **0.05**	0.95 | **0.12**	1.03 | **0.55**
	SMART (Judge: GPT-5.5)	**0.30** | **0.14**	**1.53** | **0.84**	**1.43** | **0.84**	**0.18** | **0.11**	**0.10** | **0.10**	**0.04** | **0.04**	**0.74** | **0.10**	**0.91** | **0.47**
	SMART (Claude 4.6)	**0.29** | **0.14**	**1.55** | **0.84**	**1.43** | **0.82**	**0.17** | **0.11**	**0.10** | **0.10**	**0.04** | **0.04**	0.79 | **0.10**	**0.92** | **0.47**
en→it | it→en	Gemma 3 4B	0.70 | 0.42	4.25 | 2.70	3.25 | 1.80	1.01 | 0.62	**1.69** | 0.85	0.44 | 0.15	0.62 | 0.28	2.34 | 1.39
	DeepSeek-V3.2	0.43 | 0.26	2.94 | 1.75	2.46 | 1.20	0.55 | 0.31	**1.29** | 0.77	0.20 | 0.09	**0.29** | 0.14	1.62 | 0.91
	SMART (Gemma 3 4B)	0.33 | 0.15	1.97 | 1.01	3.03 | 0.96	0.40 | 0.12	3.45 | 0.12	0.18 | 0.06	0.47 | 0.10	1.56 | 0.55
	SMART (DeepSeek-V3.2)	0.25 | 0.12	1.69 | **0.82**	2.50 | 0.76	0.33 | **0.09**	2.70 | **0.09**	**0.14** | **0.04**	0.39 | **0.08**	**1.30** | **0.44**
	SMART (Judge: GPT-5.5)	**0.23** | **0.10**	**1.45** | **0.69**	**2.19** | **0.67**	**0.29** | **0.08**	2.42 | **0.07**	**0.11** | **0.03**	**0.35** | **0.06**	**1.13** | **0.38**
	SMART (Claude 4.6)	**0.21** | **0.11**	**1.49** | **0.69**	**2.15** | **0.65**	**0.28** | **0.08**	2.37 | **0.07**	**0.11** | **0.03**	**0.35** | **0.06**	**1.13** | **0.38**

Abbreviations. Term. = Terminology; Acc. = Accuracy; Flu. = Fluency; Ling. = Linguistic Conventions; Tech. = Technical. Within each cell, the left and right values correspond to the left and right translation directions shown in the first column, respectively. SMART (Claude 4.6) is the default setting; SMART (Judge: GPT-5.5) changes only the judge model. Complete fine-grained results over all 15 bidirectional language pairs are reported in Appendix F.3: English→locale results are in Tables 15 and 16, and locale→English results are in Tables 17 and 18.

penalty by 0.41, indicating that MoA benefits from independently generated specialist hypotheses rather than combining their instructions in one prompt. Appendix F.5 defines all ablation settings.

5.5 Human Evaluation

To complement SubMQM with viewer-facing assessment, we conduct a blind human study on 20 continuous subtitle passages comprising 267 aligned lines from 13 television series. Twenty annotators, given the corresponding video and English source, rank five anonymized translations from Online, Claude Sonnet 4.6, GPT-5.5, TransAgent, and SMART on four criteria: Fidelity, Consistency, Language Quality, and Subtitle Quality. Candidate order is randomized, and ranks run from 5 (best) to 1 (worst). Instructions and the evaluation interface are in Appendix F.7.

Table 7: Human evaluation by average rank (5 best, 1 worst). Best and second-best are highlighted in blue-gray and warm beige.

Method	Fidelity	Consistency	Language	Subtitle	Overall
Online	1.10	1.10	1.05	1.10	1.05
Claude Sonnet 4.6	**3.65**	3.35	**3.80**	**3.90**	**3.75**
GPT-5.5	2.45	2.30	2.45	2.30	2.30
TransAgent	3.35	**3.70**	3.30	3.25	3.40
SMART	**4.45**	**4.55**	**4.40**	**4.45**	**4.50**

Criteria. Fidelity = meaning preservation; Consistency = terminology and cross-segment consistency; Language = fluency, naturalness, and style; Subtitle = readability and subtitle-specific presentation.

As shown in Table 7, SMART ranks first on all four criteria and achieves the highest Overall score of 4.50, with its best result on Consistency (4.55). Given the study's limited scale and coarse criteria, we use human evaluation to validate viewer-facing quality and retain SubMQM for fine-grained error analysis.

6 Conclusion

We introduced **SMART**, a self-evolving multi-agent system for long-form subtitle translation that combines persistent series-level memory, adaptive routing, tool-augmented translation, and test-time evolution without updating model parameters. We also introduced **Subtitle Arena** and **SubMQM** for evaluating long-form subtitle translation across languages and quality dimensions. Experiments show that SMART consistently improves over strong baselines, suggesting that long-form translation benefits from treating translation as a stateful and adaptive process rather than a sequence of independent sentence-level decisions.

AI USE STATEMENT

We used generative AI tools solely for language editing and polishing, including improving grammar, clarity, and readability of the manuscript. The tools were not used to generate research ideas, conduct experiments, analyze results, or make scientific claims. All AI-assisted edits were reviewed and revised by the authors. We take full responsibility for the final content of this work.

ETHICS STATEMENT

This work studies long-form subtitle translation using publicly available subtitle data. Such data may contain cultural stereotypes, offensive language, or other biases present in the original media and community-authored translations. SMART may also reproduce biases or errors from its underlying language models and external knowledge sources. We therefore evaluate the system across multiple languages and error dimensions and include human evaluation to complement automatic metrics. The system is intended to assist subtitle translation rather than replace professional judgment, particularly for culturally sensitive or high-stakes content.

REPRODUCIBILITY STATEMENT

We provide detailed descriptions of SMART's architecture, test-time training procedure, memory construction, routing policies, tool interfaces, evaluation protocol, and ablation settings in the main paper and appendix. We also document the construction and filtering of Subtitle Arena, the SubMQM evaluation rubric, human-evaluation protocol, and the model configurations used in our experiments. These details are intended to support reproduction of both the proposed system and the reported evaluation.

REFERENCES

Emre Can Acikgoz, Cheng Qian, Jonas Hübotter, Heng Ji, Dilek Hakkani-Tür, and Gokhan Tur. Tool-r0: Self-evolving llm agents for tool-learning from zero data, 2026. URL https://arxiv.org/abs/2602.21320.

Lakshya A Agrawal, Shangyin Tan, Dilara Soylu, Noah Ziems, Rishi Khare, Krista Opsahl-Ong, Arnav Singhvi, Herumb Shandilya, Michael J Ryan, Meng Jiang, et al. Gepa: Reflective prompt evolution can outperform reinforcement learning. arXiv preprint arXiv:2507.19457, 2025.

Refine AI. Subscene: A large-scale multilingual subtitle dataset, 2024.

Anthropic. Introducing claude sonnet 4.6. https://www.anthropic.com/news/claude-sonnet-4-6, 2026a.

Anthropic. Introducing claude opus 4.8. https://www.anthropic.com/news/claude-opus-4-8, 2026b.

Shiqi Chen, Jingze Gai, Ruochen Zhou, Jinghan Zhang, Tongyao Zhu, Junlong Li, Kangrui Wang, Zihan Wang, Zhengyu Chen, Klara Kaleb, Ning Miao, Siyang Gao, Cong Lu, Manling Li, Junxian He, and Yee Whye Teh. Skillcraft: Can llm agents learn to use tools skillfully?, 2026. URL https://arxiv.org/abs/2603.00718.

Yuheng Cheng, Ceyao Zhang, Zhengwen Zhang, Xiangrui Meng, Sirui Hong, Wenhao Li, Zihao Wang, Zekai Wang, Feng Yin, Junhua Zhao, and Xiuqiang He. Exploring large language model based intelligent agents: Definitions, methods, and prospects, 2024. URL https://arxiv.org/abs/2401.03428.

Chaoqun Cui, Shijing Wang, Liangbin Huang, Qingqing Gu, Zhaolong Huang, Xiao Zeng, and Wenji Mao. Hermes the polyglot: A unified framework to enhance expressiveness for multimodal interlingual subtitling. In Proceedings of the ACM Web Conference 2026, pp. 7068–7079, 2026a.

Chaoqun Cui, Shijing Wang, Liangbin Huang, Qingqing Gu, Zhaolong Huang, Xiao Zeng, and Wenji Mao. From utterance to vividity: Training expressive subtitle translation llm via adaptive local preference optimization. arXiv preprint arXiv:2602.01068, 2026b.

Mingkai Deng, Jianyu Wang, Cheng-Ping Hsieh, Yihan Wang, Han Guo, Tianmin Shu, Meng Song, Eric Xing, and Zhiting Hu. Rlprompt: Optimizing discrete text prompts with reinforcement learning. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, pp. 3369–3391, 2022.

Jinyuan Fang, Yanwen Peng, Xi Zhang, Yingxu Wang, Xinhao Yi, Guibin Zhang, Yi Xu, Bin Wu, Siwei Liu, Zihao Li, Zhaochun Ren, Nikos Aletras, Xi Wang, Han Zhou, and Zaiqiao Meng. A comprehensive survey of self-evolving ai agents: A new paradigm bridging foundation models and lifelong agentic systems, 2025. URL https://arxiv.org/abs/2508.07407.

Zhaopeng Feng, Yan Zhang, Hao Li, Bei Wu, Jiayu Liao, Wenqiang Liu, Jun Lang, Yang Feng, Jian Wu, and Zuozhu Liu. Tear: Improving llm-based machine translation with systematic self-refinement. In Findings of the Association for Computational Linguistics: NAACL 2025, pp. 3922–3938, 2025.

Patrick Fernandes, Daniel Deutsch, Mara Finkelstein, Parker Riley, André FT Martins, Graham Neubig, Ankush Garg, Jonathan H Clark, Markus Freitag, and Orhan Firat. 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, pp. 1066–1083, 2023.

Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. Promptbreeder: Self-referential self-improvement via prompt evolution. arXiv preprint arXiv:2309.16797, 2023.

Huan-ang Gao, Jiayi Geng, Wenyue Hua, Mengkang Hu, Xinzhe Juan, Hongzhang Liu, Shilong Liu, Jiahao Qiu, Xuan Qi, Yiran Wu, et al. A survey of self-evolving agents: What, when, how, and where to evolve on the path to artificial super intelligence. arXiv preprint arXiv:2507.21046, 2025.

Qingyan Guo, Rui Wang, Junliang Guo, Bei Li, Kaitao Song, Xu Tan, Guoqing Liu, Jiang Bian, and Yujiu Yang. Connecting large language models with evolutionary algorithms yields powerful prompt optimizers. arXiv preprint arXiv:2309.08532, 2023.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V. Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. In Kate Larson (ed.), Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24, pp. 8048–8057. International Joint Conferences on Artificial Intelligence Organization, 8 2024. doi: 10.24963/ijcai.2024/890. URL https://doi.org/10.24963/ijcai.2024/890. Survey Track.

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

Junyan Hu, Parijat Bhowmick, Inmo Jang, Farshad Arvin, and Alexander Lanzon. A decentralized cluster formation containment framework for multirobot systems. IEEE Transactions on Robotics, 37(6):1936–1955, 2021. doi: 10.1109/TRO.2021.3071615.

Shengran Hu, Cong Lu, and Jeff Clune. Automated design of agentic systems, 2025. URL https://arxiv.org/abs/2408.08435.

Liyan Kang, Luyang Huang, Ningxin Peng, Peihao Zhu, Zewei Sun, Shanbo Cheng, Mingxuan Wang, Degen Huang, and Jinsong Su. BigVideo: A large-scale video subtitle translation dataset for multimodal machine translation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), Findings of the Association for Computational Linguistics: ACL 2023, pp. 8456–8473, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.535. URL https://aclanthology.org/2023.findings-acl.535/.

Alina Karakanta, Matteo Negri, and Marco Turchi. Must-cinema: a speech-to-subtitles corpus. In Proceedings of the Twelfth Language Resources and Evaluation Conference, pp. 3727–3734, 2020.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T Joshi, Hanna Moazam, et al. Dspy: Compiling declarative language model calls into self-improving pipelines. arXiv preprint arXiv:2310.03714, 2023.

Tom Kocmi and Christian Federmann. Gemba-mqm: Detecting translation quality error spans with gpt-4. In Proceedings of the Eighth Conference on Machine Translation, pp. 768–775, 2023.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for "mind" exploration of large language model society. In Thirty-seventh Conference on Neural Information Processing Systems, 2023.

Weiya Li, Junjie Chen, Bei Li, Boyang Liu, Zichen Wen, Nuanqiao Shan, Xiaoqian Liu, Anping Liu, Huajie Liu, Hu Song, et al. Tactic: Translation agents with cognitive-theoretic interactive collaboration. arXiv preprint arXiv:2506.08403, 2025.

Siwei Liu, Jinyuan Fang, Han Zhou, Yingxu Wang, and Zaiqiao Meng. Sew: Self-evolving agentic workflows for automated code generation, 2026. URL https://arxiv.org/abs/2505.18646.

Yichen Lu, Wei Dai, Jiaen Liu, Ching Wing Kwok, Zongheng Wu, Xudong Xiao, Ao Sun, Sheng Fu, Jianyuan Zhan, Yian Wang, et al. Vidove: A translation agent system with multimodal context and memory-augmented reasoning. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 228–243, 2025.

Dang Nguyen, Viet Dac Lai, Seunghyun Yoon, Ryan A. Rossi, Handong Zhao, Ruiyi Zhang, Puneet Mathur, Nedim Lipka, Yu Wang, Trung Bui, Franck Dernoncourt, and Tianyi Zhou. Dynasaur: Large language agents beyond predefined actions, 2025. URL https://arxiv.org/abs/2411.01747.

Sara Papi, Marco Gaido, Alina Karakanta, Mauro Cettolo, Matteo Negri, and Marco Turchi. Direct speech translation for automatic subtitling. Transactions of the Association for Computational Linguistics, 11:1355–1376, 2023.

Ashmari Pramodya, Yusuke Sakai, Justin Vasselli, Hidetaka Kamigaito, and Taro Watanabe. Translating movie subtitles by large language models using movie-meta information. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pp. 315–330, 2025.

Reid Pryzant, Dan Iter, Jerry Li, Yin Lee, Chenguang Zhu, and Michael Zeng. Automatic prompt optimization with "gradient descent" and beam search. In EMNLP, pp. 7957–7968, 2023.

Taylor Shin, Yasaman Razeghi, Robert L Logan IV, Eric Wallace, and Sameer Singh. Autoprompt: Eliciting knowledge from language models with automatically generated prompts. In Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP), pp. 4222–4235, 2020.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. arXiv preprint arXiv:2601.03267, 2025.

Mirac Suzgun, Mert Yuksekgonul, Federico Bianchi, Dan Jurafsky, and James Zou. Dynamic cheatsheet: Test-time learning with adaptive memory, 2025. URL https://arxiv.org/abs/2504.07952.

Ye Tian, Baolin Peng, Linfeng Song, Lifeng Jin, Dian Yu, Haitao Mi, and Dong Yu. Toward self-improvement of llms via imagination, searching, and criticizing, 2024. URL https://arxiv.org/abs/2404.12253.

Jörg Tiedemann and Hengyu Luo. Opensubtitles2024: A massively parallel dataset of movie subtitles for mt development and evaluation. In Proceedings of the 15th edition of the Language Resources and Evaluation Conference (LREC 2026), 2026.

George Wang, Jiaqian Hu, and Safinah Ali. Maats: A multi-agent automated translation system based on mqm evaluation. arXiv preprint arXiv:2505.14848, 2025a.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models, 2023. URL https://arxiv.org/abs/2305.16291.

Jiaan Wang, Fandong Meng, Yunlong Liang, and Jie Zhou. Drt: Deep reasoning translation via long chain-of-thought. In Findings of the Association for Computational Linguistics: ACL 2025, pp. 6770–6782, 2025b.

Yingxu Wang, Siwei Liu, Jinyuan Fang, and Zaiqiao Meng. Evoagentx: An automated framework for evolving agentic workflows, 2025c. URL https://arxiv.org/abs/2507.03616.

Yutong Wang, Jiali Zeng, Xuebo Liu, Derek F. Wong, Fandong Meng, Jie Zhou, and Min Zhang. Delta: An online document-level translation agent based on multi-level memory, 2025d. URL https://arxiv.org/abs/2410.08143.

Tianxin Wei, Noveen Sachdeva, Benjamin Coleman, Zhankui He, Yuanchen Bei, Xuying Ning, Mengting Ai, Yunzhe Li, Jingrui He, Ed H. Chi, Chi Wang, Shuo Chen, Fernando Pereira, Wang-Cheng Kang, and Derek Zhiyuan Cheng. Evo-memory: Benchmarking llm agent test-time learning with self-evolving memory, 2025. URL https://arxiv.org/abs/2511.20857.

Patrick Wilken, Panayota Georgakopoulou, and Evgeny Matusov. SubER - a metric for automatic evaluation of subtitle quality. In Elizabeth Salesky, Marcello Federico, and Marta Costa-jussà (eds.), Proceedings of the 19th International Conference on Spoken Language Translation (IWSLT 2022), pp. 1–10, Dublin, Ireland (in-person and online), May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.iwslt-1.1. URL https://aclanthology.org/2022.iwslt-1.1/.

Minghao Wu, Jiahao Xu, and Longyue Wang. TransAgents: Build your translation company with language agents. In Delia Irazu Hernandez Farias, Tom Hope, and Manling Li (eds.), Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 131–141, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-demo.14. URL https://aclanthology.org/2024.emnlp-demo.14/.

Minghao Wu, Jiahao Xu, Yulin Yuan, Gholamreza Haffari, Longyue Wan, Weihua Luo, and Kaifu Zhang. (perhaps) beyond human translation: Harnessing multi-agent collaboration for translating ultra-long literary texts. Transactions of the Association for Computational Linguistics, 13:901–922, 2025.

Haoran Xu, Young Jin Kim, Amr Mohamed Nabil Aly Aly Sharaf, and Hany Awadalla. A paradigm shift in machine translation: Boosting translation performance of large language models. In International Conference on Learning Representations, volume 2024, pp. 2747–2767, 2024.

Haoran Xu, Kenton Murray, Philipp Koehn, Hieu Hoang, Akiko Eriguchi, and Huda Khayrallah. X-alma: Plug & play modules and adaptive rejection for quality translation at scale. In International Conference on Learning Representations, volume 2025, pp. 94654–94680, 2025a.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. arXiv preprint arXiv:2509.17765, 2025b.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. arXiv preprint arXiv:2505.09388, 2025.

Yaoning Yu, Ye Yu, Peiyan Zhang, Kai Wei, Haojing Luo, and Haohan Wang. Sipdo: Closed-loop prompt optimization via synthetic data feedback. arXiv preprint arXiv:2505.19514, 2025.

Mert Yuksekgonul, Federico Bianchi, Joseph Boen, Sheng Liu, Zhi Huang, Carlos Guestrin, and James Zou. Textgrad: Automatic" differentiation" via text. arXiv preprint arXiv:2406.07496, 2024.

Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman. Star: Bootstrapping reasoning with reasoning, 2022. URL https://arxiv.org/abs/2203.14465.

Weihao Zeng, Yuzhen Huang, Lulu Zhao, Yijun Wang, Zifei Shan, and Junxian He. B-star: Monitoring and balancing exploration and exploitation in self-taught reasoners, 2025. URL https://arxiv.org/abs/2412.17256.

Guibin Zhang, Kaijie Chen, Guancheng Wan, Heng Chang, Hong Cheng, Kun Wang, Shuyue Hu, and Lei Bai. Evoflow: Evolving diverse agentic workflows on the fly, 2025a. URL https://arxiv.org/abs/2502.07373.

Haozhen Zhang, Quanyu Long, Jianzhu Bao, Tao Feng, Weizhi Zhang, Haodong Yue, and Wenya Wang. Memskill: Learning and evolving memory skills for self-evolving agents, 2026. URL https://arxiv.org/abs/2602.02474.

Jiayi Zhang, Jinyu Xiang, Zhaoyang Yu, Fengwei Teng, Xionghui Chen, Jiaqi Chen, Mingchen Zhuge, Xin Cheng, Sirui Hong, Jinlin Wang, Bingnan Zheng, Bang Liu, Yuyu Luo, and Chenglin Wu. Aflow: Automating agentic workflow generation, 2025b. URL https://arxiv.org/abs/2410.10762.

Peiyan Zhang, Haibo Jin, Leyang Hu, Xinnuo Li, Liying Kang, Man Luo, Yangqiu Song, and Haohan Wang. Revolve: Optimizing ai systems by tracking response evolution in textual optimization. arXiv preprint arXiv:2412.03092, 2024.

Andrew Zhao, Daniel Huang, Quentin Xu, Matthieu Lin, Yong-Jin Liu, and Gao Huang. Expel: Llm agents are experiential learners, 2024. URL https://arxiv.org/abs/2308.10144.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. Advances in neural information processing systems, 36:46595–46623, 2023.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. Large language models are human-level prompt engineers. In The eleventh international conference on learning representations, 2022.

A Extended Related Work

Multi-Agent Systems and Translation Agents. Multi-agent LLM systems distribute reasoning across agents with specialized roles, enabling decomposition, cross-agent verification, and iterative refinement (Li et al., 2023; Hu et al., 2021; Guo et al., 2024; Cheng et al., 2024). In machine translation, existing approaches can broadly be grouped into role-based collaboration and context-aware translation. MAPS (He et al., 2024), TransAgents (Wu et al., 2024), TEaR (Feng et al., 2025), and DRT (Wang et al., 2025b) decompose translation into complementary processes such as knowledge elicitation, generation, estimation, criticism, and refinement. TACTIC (Li et al., 2025) and MAATS (Wang et al., 2025a) further specialize agents for contextual reasoning, external knowledge, and individual MQM error categories. For long-context translation, DelTA (Wang et al., 2025d) maintains multi-level memory for document consistency, while ViDove (Lu et al., 2025) incorporates visual, auditory, and historical context for video subtitle translation. ALPO (Cui et al., 2026b) takes a different direction and improves expressive subtitle translation through adaptive preference optimization. Although these methods substantially improve translation quality and contextual consistency, their workflows, agent roles, or learned policies are generally fixed once deployed.

Self-Evolving LLM Agents. Self-evolving agents seek to improve agent behavior through continued interaction. One line of work evolves the structure and execution flow of agent systems, including ADAS (Hu et al., 2025), AFlow (Zhang et al., 2025b), EvoAgentX (Wang et al., 2025c), EvoFlow (Zhang et al., 2025a), and SEW (Liu et al., 2026). Other methods accumulate reusable skills, tools, or experience (Wang et al., 2023; Zhao et al., 2024; Nguyen et al., 2025; Zhang et al., 2026; Acikgoz et al., 2026; Chen et al., 2026), improve model reasoning through self-generated supervision (Zelikman et al., 2022; Zeng et al., 2025; Tian et al., 2024), or evolve persistent memory at inference time (Suzgun et al., 2025; Wei et al., 2025). Together, these approaches establish workflows, capabilities, models, and memory as complementary targets for continual adaptation (Gao et al., 2025; Fang et al., 2025). However, prior work mainly evaluates whether adaptation improves subsequent tasks, with less attention to whether newly acquired behaviors interfere with previously learned ones. For long-form translation, this distinction is important because improvements on later sentences should not destabilize terminology, register, or translation conventions established earlier.

Prompt Optimization. Automated prompt optimization adapts LLM behavior without updating model parameters. Earlier methods formulate prompt construction as discrete search or reinforcement learning (Shin et al., 2020; Deng et al., 2022), while later approaches use LLMs themselves to generate, mutate, and select prompts (Zhou et al., 2022; Guo et al., 2023; Fernando et al., 2023; Khattab et al., 2023). Feedback-driven methods are particularly related to SMART. APO (Pryzant et al., 2023) interprets critiques as textual gradients, TextGrad (Yuksekgonul et al., 2024) propagates textual feedback through computational graphs, and REVOLVE (Zhang et al., 2024), SIPDO (Yu et al., 2025), and GEPA (Agrawal et al., 2025) iteratively refine prompts through reflection and generated feedback. These methods demonstrate that natural-language feedback can effectively improve instructions, but typically optimize a prompt against a fixed task or objective. SMART instead maintains agent instructions as persistent state that evolves throughout an ongoing translation process.

LLM-as-a-Judge. LLMs have increasingly been adopted as scalable evaluators of model-generated outputs. General LLM-as-a-judge methods use strong language models to score, rank, or critique responses, while also exposing evaluation biases such as position or verbosity preferences (Zheng et al., 2023). This paradigm is especially useful for machine translation, where translation errors cannot always be summarized by a single lexical similarity score. GEMBA-MQM (Kocmi & Federmann, 2023) and AutoMQM (Fernandes et al., 2023) use LLMs to provide structured MQM-style judgments over error spans, categories, and severities. Such feedback offers richer diagnostic information than scalar quality scores and can support targeted correction. Existing work, however, primarily treats the judge as an evaluation endpoint or as feedback for local refinement. SMART additionally converts translation-specific judgments into persistent updates of the translation system, coupling evaluation with continual test-time training.

B SMART IMPLEMENTATION DETAILS

B.1 SUBTITLE PARSING AND CONTEXT CONSTRUCTION

Each SRT block is parsed into its index, start and end timestamps, and source text. SMART augments the current block with neighboring source lines, scene metadata, and the series content profile. Consecutive blocks that appear to form a single interrupted sentence are grouped into one sentence and are split back to their original timestamp boundaries after translation. This avoids forcing the model to translate each block independently when its semantics are sentence-level.

B.2 MEMORY CONSTRUCTION AND CONTENT PROFILING

The terminology pool stores source terms together with confirmed target renderings and occurrence statistics. Character memory contains names, target-language renderings when confidently established, roles. Domain memory stores series- and scene-level facts required to interpret technical language, cultural references, and ambiguous slang. The idiom bank stores target-language expressions indexed by communicative situation and genre.

At the beginning of a series, SMART first builds an initial profile of genre, setting, principal characters, and likely domain terminology. During translation, uncertain expressions can trigger targeted retrieval by web search. Search results are passed through a contextual interpretation step before being exposed to translators, so retrieved literal meanings are not automatically treated as appropriate subtitle renderings. When a series already has persistent memory, existing information is reused and only missing entries are queried.

B.3 ROUTING POLICY AND TRANSLATOR ROLES

The router uses features including source length, subtitle duration, scene tone, emotional markers, slang/profanity indicators, and display-budget pressure. Inputs are mapped to a small number of routing categories, and each category activates a subset of the translator pool. The default path uses the faithful and natural translators; length-constrained sentences may additionally activate the length-aware translator, emotionally marked dialogue may activate the expressive translator, and slang-heavy dialogue may activate the colloquial translator. During training, routing outcomes are logged together with judge scores and become evidence for updating ρ_t.

The translator roles are:

- **Faithful translator**: prioritizes semantic accuracy and preservation of established entities and domain terms.
- **Natural translator**: prioritizes target-language fluency and register-appropriate phrasing.
- **Expressive translator**: prioritizes character narrative voice, emotion, and dramatic force when these are salient.
- **Length-aware translator**: explicitly conditions on subtitle duration and display budgets.
- **Colloquial translator**: handles slang, profanity, and conversational language without over-literal or over-sanitized rendering.

B.4 TOOL INTERFACES

SMART exposes eight callable tools. `terminology_lookup` retrieves established renderings, while `terminology_register` proposes new entries with conflict detection. `memory_search` retrieves similar prior subtitle units and their translations. `get_context` returns surrounding dialogue, existing translations, scene information, and domain notes. `constraint_check` measures line length and reading-speed violations. `web_search` retrieves external evidence for jargon, cultural references, and unclear slang. `idiom_lookup` returns situation-appropriate target-language idioms from the series bank or a fallback query. `fluency_check` flags translationese, over-colloquialization, and register mismatch. Tool calls and outputs are recorded during training and can inform subsequent routing updates.

Table 8 lists the purpose and execution of each. All tools run locally except `web_search`, which issues an external query and passes the raw result through the contextual interpretation step of

Table 8: Tool modules available to SMART translators. Execution is local unless otherwise noted.

Tool	Purpose	Execution
`terminology_lookup`	Retrieve the established target rendering of a recurring name or term.	Match against the shared terminology table and return the confirmed translation or partial matches.
`terminology_register`	Fix the target rendering of a new name or term.	Write to the terminology table with conflict detection to prevent overwriting an existing entry.
`memory_search`	Find previously translated sentences with similar source content.	Rank entries from the current episode's translation memory by source overlap to support consistency and style.
`get_context`	Retrieve surrounding dialogue, scene information, and domain notes.	Read up to eight preceding and succeeding sentences together with the associated scene metadata.
`constraint_check`	Verify subtitle reading-speed and line-length constraints.	Compute characters per second and per-line character counts, and report specific violations.
`web_search`	Resolve jargon, cultural references, or ambiguous slang.	External search whose raw results are interpreted to separate literal meaning from the register appropriate to the scene.
`idiom_lookup`	Retrieve target-language idioms and colloquial expressions.	Query a content-specific idiom bank, with a dynamic lookup when the requested category is unavailable.
`fluency_check`	Detect translationese, excessive colloquialization, and register mismatch.	Evaluate the target-language draft and return localized revision suggestions.

Appendix C.2 before returning it. The router fixes which subset of tools each activated translator may call, and this assignment is itself updated during self-evolution (Appendix C.6).

B.5 Judge–Refiner and Rescue Step

The judge evaluates candidate translations using semantic accuracy, fluency, register, expressiveness when relevant, terminology consistency, and subtitle constraints, and reports an overall score on a 0–10 scale. The top-scoring candidate is passed to the refiner together with its critique. If that score falls below the rescue threshold $\theta = 8.0$, SMART performs a rescue pass by re-running the MoA translator with the judge critique appended and comparing the new output against the existing candidates. The accepted translation updates the local translation history and, when appropriate, the persistent series memory.

B.6 Episode-Level Consistency Pass

After sentence-level translation, SMART performs an episode-level consistency pass over overlapping windows of increasing size. Each window is conditioned on locked preceding translations to reduce drift. The editor checks recurring terminology, dialogue coherence, pronoun and reference consistency, and residual translationese. Later rounds are skipped when an earlier round makes no changes.

B.7 Test-Time Training Loop

Each series is partitioned chronologically using a 3:7 training/inference ratio. Prompt and routing updates are allowed only on the training prefix. The evolved prompts and router are frozen on the held-out portion, while translation memory and series memory continue to update because they represent task state rather than learned model parameters.

The self-evolution procedure described in the main text proceeds over epochs of the training subset $\mathcal{D}_{\text{tr}}$. Each epoch runs the same per-sentence pipeline as inference, then evaluates the collected scores

and updates the prompts and routing policy. The learnable state is the set of agent prompts, the routing policy, and the tool assignments, all expressed as text or configuration rather than model weights. The five phases of one epoch are as follows.

```
Phase 1  TRANSLATE
  translate the training files with the current config
  collect per segment: judge score, agent used, tools called, category

Phase 2  EVALUATE
  aggregate into statistics:
    per agent    : mean/min/max score, tool usage counts
    per category : which agent scores highest
    worst cases  : lowest-scoring sentences with critiques

Phase 3  PROMPT OPTIMIZATION  (F^Pi)
  for each agent with mean score below a threshold:
    feed current prompt + statistics + worst cases to the evaluator
    the evaluator rewrites the prompt to address the failure modes

Phase 4  STRUCTURE OPTIMIZATION  (F^rho)
  feed routing policy + per-category performance to the evaluator
  update: drop underperforming agents from a category,
          reassign tools, keep at least two agents per category

Phase 5  SAVE CONFIG
  persist evolved prompts, routing policy, tool assignments, stats
  this config is loaded unchanged at inference time
```

The loop terminates after a fixed number of epochs, and the configuration from the final epoch is the (Π_T, ρ_T) applied during test-time inference. Because every update is a text rewrite driven by aggregated judge feedback, the procedure requires no gradient computation and leaves the underlying LLM unchanged.

B.8 Instantiation of F^{Π} and F^{ρ}

Both operators are single-proposal rewrites over constrained spaces, applied once per training epoch from statistics aggregated over the entire training prefix. The series memory M is initialized once for each series before the first training epoch and is carried across epochs, so prompt and routing updates build on the same persistent task state rather than reconstructing series knowledge from scratch. Algorithm 1 states the full loop; the paragraphs below give the parts that the pseudocode abbreviates.

F^{Π}: prompt rewrite. The evaluator uses the aggregated per-agent statistics and low-scoring cases from Phase 2 to identify agents with recurring failure modes; prompts for agents not flagged by this feedback are left unchanged. For each flagged agent, the operator receives the current prompt, its aggregate statistics, and representative low-scoring sentences together with their critiques, and produces one rewritten prompt. The rewrite is constrained: it must preserve the agent's role statement and the workflow skeleton, and may add or amend rules but may not redefine the role or delete a workflow step. The critiques describe failure modes rather than supplying corrected outputs (Appendix C.5), so the rewrite cannot encode sentence-specific fixes.

F^{ρ}: table rewrite. Its input is the current routing table, the per-category record of which agents achieved the highest judge scores, and tool-usage counts; its output is a new table. Two constraints bound it: every category retains at least two agents, and the category set is fixed, so the operator reassigns agents and tools within a fixed schema rather than inventing categories. The reachable space is therefore finite and enumerable.

A proposal satisfying the structural constraints above is adopted; there is no separate validation-gated acceptance test。 The loop runs a fixed number of epochs and the final epoch's configuration becomes (Π_T, ρ_T). We run self-evolution for three epochs over $\mathcal{D}_{\mathrm{tr}}$. After the third epoch, the resulting prompts, routing policy, and tool assignments are frozen and used for test-time inference on $\mathcal{D}_{\mathrm{inf}}$.

Algorithm 1 Self-evolution at test-time training for one series. Lines 19–24 are F^{Π}; line 27 is F^{ρ}. Both act once per epoch on statistics aggregated over the whole training prefix.

Require: training prefix $\mathcal{D}_{\text{tr}}$ (first 30% of the series); translator pool $\mathcal{R}$; tool set $\mathcal{T}$; judge J; refiner R; evaluator E
Require: initial prompts Π_0, initial routing table ρ_0, epochs T, agent threshold τ_{agent}, rescue threshold θ, worst-case budget k
Ensure: frozen configuration (Π_T, ρ_T)
1: $M \leftarrow \emptyset$ ▷ series memory
2: **for** $t = 0$ **to** $T - 1$ **do**
3: $\mathcal{L} \leftarrow \emptyset$ ▷ Phase 1: translate
4: **for** each episode $e \in \mathcal{D}_{\text{tr}}$ **do**
5: **for** each sentence $d_i \in e$ **do**
6: $\mathcal{A}_i \leftarrow \rho_t(\phi(d_i, M))$ ▷ category lookup, no learned selector
7: $y_i^{(r)} \leftarrow r(d_i, \pi_{r,t}, M, \mathcal{T})$ for each $r \in \mathcal{A}_i$
8: $\left(c_i^{(r)}, \delta_i^{(r)}\right) \leftarrow J\left(y_i^{(r)}, d_i, M\right)$ for each $r \in \mathcal{A}_i$
9: $r^\star \leftarrow \arg\max_{r \in \mathcal{A}_i} c_i^{(r)}$
10: **if** $c_i^{(r^\star)} < \theta$ **then**
11: rerun the faithful translator with $\delta_i^{(r^\star)}$ appended; keep the better candidate
12: **end if**
13: $\hat{y}_i \leftarrow R\left(y_i^{(r^\star)}, \delta_i^{(r^\star)}, M\right)$; $M \leftarrow \textsc{UpdateMemory}(M, \hat{y}_i)$
14: append $\left(\text{category}(d_i), \mathcal{A}_i, r^\star, \{c_i^{(r)}, \delta_i^{(r)}\}, \text{tools called}\right)$ to $\mathcal{L}$
15: **end for**
16: episode-level consistency pass over overlapping windows
17: **end for**
18: $(S_{\text{agent}}, S_{\text{cat}}, W) \leftarrow E(\mathcal{L})$ ▷ Phase 2: Δ_t; W holds k worst sentences per agent
19: **for all** $r \in \mathcal{R}$ **do** ▷ Phase 3: F^{Π}
20: **if** $\text{mean}(S_{\text{agent}}[r]) < \tau_{\text{agent}}$ **then**
21: $\pi_{r,t+1} \leftarrow \textsc{Rewrite}(\pi_{r,t}, S_{\text{agent}}[r], W_r)$
22: **s.t.** role statement and workflow steps preserved
23: **else**
24: $\pi_{r,t+1} \leftarrow \pi_{r,t}$ ▷ at or above threshold: untouched
25: **end if**
26: **end for**
27: $\rho_{t+1} \leftarrow \textsc{Rewrite}(\rho_t, S_{\text{cat}})$ ▷ Phase 4: F^{ρ}
28: **s.t.** $|\rho_{t+1}(c)| \geq 2$ for every category c; category set fixed
29: persist $(\Pi_{t+1}, \rho_{t+1}, S_{\text{agent}}, S_{\text{cat}})$ ▷ Phase 5
30: **end for**
31: **return** (Π_T, ρ_T) ▷ final epoch; no acceptance test, rollback, or early stopping

C Worked Examples

The examples below trace SMART on episodes of the crime drama Bosch, translating from English to Chinese. They are illustrative of one language direction; the same mechanisms apply to every target language in **Subtitle Arena**. We show a full sentence traversal of the per-sentence pipeline, then isolate the two components, the web-search contextual retrieval path and the cross-episode series memory.

C.1 End-to-End Sentence Example

Consider a sentence from a scene in which one detective needles another about a stalled case. The router extracts surface features, classifies the sentence, and activates the corresponding subset of translators; each activated translator runs its own tool loop before emitting a candidate; the judge scores all candidates and the refiner finalizes the output.

```
Input sentence
  index    : 128
  time     : 00:12:04,320 --> 00:12:06,880   (duration 2.56s)
  source   : Bosch, you're chasing your own tail here.

Step 1  Routing
  features : {word_count: 8, duration: 2.56, has_slang: false,
              has_emotion: false, scene_tone: "tense"}
  category : expressive          (scene tone is tense)
  agents   : [faithful, natural, expressive]

Step 2  Agent execution (each agent runs its own tool loop)
  [faithful]
    tool_use  terminology_lookup {"term": "Bosch"}
    result    {found: true, target: 博斯, category: name}
    tool_use  constraint_check {"text": 博斯，你这是在原地打转。,
                                "duration": 2.56}
    result    {valid: true, cjk_chars: 11, max_chars: 38, cps: 4.3}
    output    博斯，你这是在原地打转。
  [natural]
    output    博斯，你这么查是在白费劲。
  [expressive]
    tool_use  idiom_lookup {"category": "futile_effort"}
    result    {idioms: [徒劳无功, 缘木求鱼, 竹篮打水], ...}
    output    博斯，你这是缘木求鱼。

Step 3  Judge (scores abbreviated)
  1 [faithful]   accuracy 9 fluency 8 register 8 ... overall 8.2
  2 [natural]    accuracy 9 fluency 9 register 9 ... overall 8.8
  3 [expressive] accuracy 7 fluency 9 register 7 ... overall 7.9
     critique(3): idiom overstates the register for a terse jab
  winner : [natural] 8.8

Step 4  Refiner + state update
  score 8.8 >= threshold, output kept as is
  final  : 博斯，你这么查是在白费劲。
  memory : translation memory += (source, final)
           terminology "Bosch"->"博斯" count += 1
```

The example shows the two properties the router is designed for: only the relevant translators are activated (the colloquial and length-aware translators are not needed here), and the terminology tool supplies the confirmed rendering of the recurring name so that the candidate does not re-derive it.

C.2 Web Search and Contextual Retrieval Example

Police procedurals contain jargon whose literal reading differs from its conversational intent. When a translator encounters such a term, it calls `web_search`; the raw result is passed through a contextual interpretation step that separates the literal domain meaning from the register the scene calls for, and the outcome is written into the scene domain notes so that later sentences and the judge see the same context.

```
Segment
  source : We got a 187 in Hollywood, roll on it.
  scene  : squad room, tone = urgent

web_search {"query": "What does 187 mean in US police radio code?",
            "context": "detective assigning a case, urgent tone"}

raw result
  California Penal Code section 187 defines the crime of murder;
  "a 187" is police radio shorthand for a homicide.

=== translation guidance ===
  literal : 187 is the penal-code number for homicide/murder.
  context : squad-room dispatch, urgent; use the natural term a
            Chinese audience expects, not the code number.
  advice  : render as 凶杀案/命案, not the literal number "187".

scene.domain_notes += "[187] homicide radio code; render as 凶杀案"

final translation
  好莱坞发生一起凶杀案，马上去。
```

Without retrieval, a literal system emits the number 187, which is meaningless to the audience. The contextual interpretation step is what prevents the opposite failure of injecting the technical penal-code register into casual speech; it reports the literal meaning and the register separately, and the domain note it writes is reused by every subsequent sentence in the scene.

C.3 Series Memory Example

The series memory persists terminology, character profiles, domain facts, and the idiom bank across episodes of the same series. A name fixed in an early episode is retrieved directly in later episodes, however far apart the occurrences are, which is the mechanism behind the cross-episode consistency the benchmark measures.

```
Episode S01E01  (memory starts empty)
  content profiling registers principal characters:
    "Harry Bosch" -> 哈里·博斯
    "J. Edgar"    -> 杰瑞·埃德加
  domain facts: LAPD Hollywood Division homicide unit
  at episode end, series memory is persisted:
    terminology : "Harry Bosch" -> {target: 哈里·博斯, count: 14}
                  "J. Edgar"    -> {target: 杰瑞·埃德加, count: 6}
    characters  : Harry Bosch (detective), J. Edgar, ...
    idiom_bank  : crime_investigation, interrogation, ...

Episode S01E05  (series memory loaded before translation)
  inject_into_executor : 23 terms preloaded into terminology table
  inject_into_profile  : characters, domain facts, idiom bank reused
                         (idiom bank rebuild skipped)

  segment source : Edgar, get Bosch on the phone.
    terminology_lookup "Bosch"  -> found, target = 博斯
    terminology_lookup "Edgar"  -> found, target = 埃德加
    final : 埃德加，给博斯打电话。

  merge back at episode end:
    new terms confirmed this episode are added to memory
    counts of existing terms accumulated (Harry Bosch: 14 -> 27)
```

Because the mapping `Harry Bosch` → 哈里·博斯 is loaded from memory rather than re-inferred, the fifth episode reuses the exact rendering fixed in the first, and the accumulated occurrence counts let the judge treat the established translation as the consistency reference for the remainder of the series.

C.4 AGENT REASONING TRACE

The examples above summarize which tools were called; here we show the interleaved reasoning that produces those calls. Each translator follows a reason-act loop: it states what it is uncertain about, calls a tool to resolve that uncertainty, reads the result, and decides the next step, until it commits to a translation. The trace below is from the natural translator on a sentence whose difficulty is not the vocabulary but the register: a line that reads as a factual statement but is meant as a sarcastic jab.

```
Segment
  source   : Great, another dead end. Love this job.
  scene    : stakeout gone wrong, tone = frustrated/sarcastic

Agent: natural  (reasoning interleaved with tool calls)

  thought : Two sentences. "dead end" is an investigation term
            but here it is figurative. "Love this job" is not
            literal; the scene tone is frustrated, so it is
            sarcasm. A literal 我爱这份工作 would read as sincere
            and invert the meaning. I first confirm the tone.
  action  : get_context {"direction": "before", "count": 2}
  result  : prev lines show the lead just fell through;
            scene.tone = "frustrated/sarcastic"

  thought : Tone confirmed as sarcastic. I should not translate
            "Love this job" straight. I want a Chinese phrasing
            that carries the same dry sarcasm rather than a
            literal statement. Check for a fitting expression.
  action  : idiom_lookup {"category": "sarcasm/complaint"}
  result  : {colloquial: [这活儿真"好", 又白跑一趟, 干这行真"值"], ...}

  thought : "又白跑一趟" matches "another dead end" naturally.
            For "Love this job" I keep the sarcasm with a dry
            这工作真"带劲", the quotes marking irony. Now verify
            it reads naturally and is not over-colloquial.
  action  : fluency_check {"text": "又白跑一趟。这工作真"带劲"。",
                           "tone": "sarcastic"}
  result  : {fluent: true, register: ok,
             note: irony quotes read well, keep as is}

  thought : Fluency check passes. Length is short, no constraint
            risk. Commit.
  output  : 又白跑一趟。这工作真"带劲"。
```

The trace shows why the reasoning matters: the hard decision is not lexical but pragmatic, and the agent reaches the sarcastic reading by first confirming the scene tone through `get_context`, then choosing an ironic phrasing over the literal one and validating its register with `fluency_check`. A system that translated the surface form would emit a sincere 我爱这份工作, reversing the speaker's intent. This interleaved reason-act record is also what the judge and the self-evolution stage consume: the recorded thoughts and tool results make it possible to attribute a low score to a specific decision rather than to the sentence as a whole.

C.5 PROMPT UPDATE EXAMPLE

During test-time training the evaluator aggregates judge scores over a batch, identifies recurring deficiencies attributable to a specific agent, and rewrites that agent's prompt through the operator F^{Π} defined in the main text. The update is expressed entirely in natural language; no LLM weights change. The example below shows the faithful agent's prompt before and after one such update, together with the failure pattern that drove it.

```
Prompt at step t  (faithful agent, expert)
  PRIORITY: semantic accuracy; preserve exact meaning.
  WORKFLOW:
   1. terminology_lookup for names/terms
   2. translate preserving full meaning
   3. constraint_check with translation and duration
   4. terminology_register for new names

Failure pattern found by the evaluator over the batch
   - interjections/onomatopoeia translated literally
     ("Oh, boy." -> 哦，孩子。),  mean score 4.2
   - short ambiguous lines translated without get_context
   - emotional register flattened in colloquial sentences

Textual critique delta_t (fed to F^Pi)
   "add rules for interjections, short-line context, and
    register preservation; keep the accuracy priority."

Prompt at step t+1  (after F^Pi, added rules)
  KEY RULES (new):
   - interjections (oh, huh, ugh): translate the FUNCTION,
     not the literal sound
   - short lines (<= 5 chars): call get_context FIRST
   - profanity/slang: match INTENSITY, do not sanitize
  WORKFLOW: (unchanged steps, get_context added at step 1)

Same input, before and after the update
  source : Oh, boy. Here we go again.   (scene tone = weary)
  before : 哦，孩子。我们又来了。            judge 4.1
  after  : 唉，又来了。                     judge 8.6
```

The critique names the failure modes rather than supplying corrected outputs, so the rewritten prompt generalizes to unseen sentences with the same difficulty rather than memorizing specific fixes. The before/after pair is a sentence the updated prompt was never shown: the literal reading of the interjection and the redundant subject both disappear once the interjection rule is in force.

C.6 Routing Policy Update Example

In parallel with the prompt update, the router policy is updated through F^{ρ} from the per-category statistics of which agents achieved the highest judge scores. The example shows the policy and one agent's tool assignment before and after an update driven by the recorded scores.

```
Routing policy at step t
  colloquial : [faithful, natural, colloquial]
  expressive : [faithful, natural, expressive]
  short      : [faithful, natural]
  long       : [faithful, natural, length_aware]
  default    : [faithful, natural]

Per-category statistics over the batch
  - faithful averages 6.2 on "colloquial" (too literal for slang)
  - natural averages 8.1 on "expressive", above expressive (7.9)

Routing policy at step t+1  (after F^rho)
  colloquial : [natural, colloquial]     (faithful dropped)
  expressive : [faithful, natural, expressive]
  short      : [faithful, expressive]    (emotion helps short lines)
  long       : [faithful, natural, length_aware]
  default    : [faithful, natural]

Tool assignment update (colloquial agent)
  + memory_search     (reuse prior slang renderings)
  + web_search        (resolve unfamiliar slang meaning)

Same input, before and after the update
  source : He's been jerking us around all day.  (category: colloquial)
  before : [faithful]   他一整天都在让我们绕圈子。     judge 6.2
  after  : [colloquial] 他耍了我们一整天。            judge 8.7
```

The policy update removes an agent from a category where it consistently underperformed and promotes one that scored higher, while the tool-assignment update biases each agent toward the tools that helped on similar inputs. Both changes are recorded so that the evolved configuration (Π_T, ρ_T) can be applied unchanged during inference. In the before/after pair, the winning candidate changes because the category no longer activates the agent that scored 6.2 on it, not because any agent was retrained.

D Subtitle Arena: Additional Details

D.1 Source Corpus and Episode Reconstruction

Subtitle Arena is derived from the OpenSubtitles collection (Tiedemann & Luo, 2026). For each target locale we read the English–Target sentence alignments released with that collection, so English is the pivot side of every bitext; both translation directions are then obtained by exchanging the roles of the two sides rather than by collecting a separate corpus for each direction. Each subtitle document is associated with an IMDb identifier and season/episode indices. We use the tuple $(\mathrm{IMDb}, \mathrm{season}, \mathrm{episode})$ as the canonical episode key and group the English and target-language versions of the same television episode under this identifier. This reconstruction converts sentence-level parallel data into episode-level subtitle streams while preserving the original narrative boundary.

D.2 SRT Reconstruction and Filtering

For each aligned sentence pair we recover the timestamp span and surface form on both sides from their respective subtitle documents and emit standard SRT blocks in presentation order, so that either side can serve as the source stream. Timestamp strings are normalized to the canonical `hh:mm:ss,ms` format. Empty or one-sided alignment groups are discarded, as are episodes whose target subtitle stream is missing or cannot be parsed.

Subtitle files may divide a single sentence across multiple adjacent display blocks. We therefore merge continuation blocks when reconstructing aligned sentences while retaining the corresponding temporal boundaries in the episode representation.

Table 9: Subtitle Arena against existing subtitle translation resources. Unit is the largest span the resource is organized around; Series context indicates whether context persists across episodes of the same title; Display constraints indicates whether subtitle timing and line-length limits are part of the evaluation.

Resource	Unit	Domain	Directions	Series context	Display constraints	Evaluation
MuST-Cinema	Talk	TED talks (speech)	en→7 languages	×	✓	BLEU
BigVideo	Clip	Web video	en→zh	×	×	BLEU
OpenSubtitles2024	Sentence	Film and television	Many	×	×	Corpus only
SubScene	Sentence	Film and television	Many	×	×	Corpus only
MuSC	Utterance segment	Streaming programs	6 directions	×	×	Accuracy, naturalness, vividness
Subtitle Arena	**Series**	**Television**	**30 directions** (15 locales, both ways)	✓	✓	**SubMQM** (7 dimensions, 19 error types)

D.3 Locale Normalization

Subtitle Arena reports 15 target locales: Chinese (zh_CN), Korean (ko_KR), European and Brazilian Portuguese (pt_PT, pt_BR), Italian (it_IT), Norwegian (no_NO), Danish (da_DK), Dutch (nl_NL), Romanian (ro_RO), Swedish (sv_SE), Turkish (tr_TR), European and Mexican Spanish (es_ES, es_MX), German (de_DE), and French (fr_FR). OpenSubtitles does not expose a distinct subtitle stream for every regional variety represented in our evaluation.

D.4 Benchmark Statistics and Translation Directions

The final benchmark contains 192 television series, 6,267 English source episodes, and 70,664 aligned bilingual episode pairs. Individual series contain between 2 and 198 episodes, with an average of 32.6 episodes per series, and span production years from 1959 to 2023. The benchmark covers 14 genres; Appendix D.6 visualizes per-locale episode coverage and genre diversity.

Because every episode pair aligns English with one target locale, the 15 locales yield 30 evaluation directions: 15 English→locale and 15 locale→English. We evaluate all of them, and report English→locale results in Appendix F.1 and locale→English results in Appendix F.2.

D.5 Relation to Existing Subtitle Benchmarks

Subtitle Arena complements prior subtitle and multimodal translation resources rather than replacing them. MuST-Cinema (Karakanta et al., 2020) targets speech-to-subtitle translation of TED talks, while BigVideo (Kang et al., 2023) provides video–subtitle supervision for multimodal MT. OpenSubtitles2024 (Tiedemann & Luo, 2026) offers broad multilingual parallel subtitle data at the sentence level, and SubScene (AI, 2024) further broadens multilingual subtitle coverage. Closest to our setting is MuSC (Cui et al., 2026b), a multidirectional subtitle corpus built from streaming programs and used to train and evaluate expressive subtitle translation; it is organized as multi-line utterance sentences rather than as complete episodes, and its evaluation targets vividness at the segment level. Table 9 summarizes the comparison. In contrast to all of these, Subtitle Arena organizes bilingual subtitle data explicitly around complete episodes and television series, making persistent discourse context and long-horizon consistency first-class evaluation targets, and pairs them with an error-typed subtitle evaluation protocol (Appendix E).

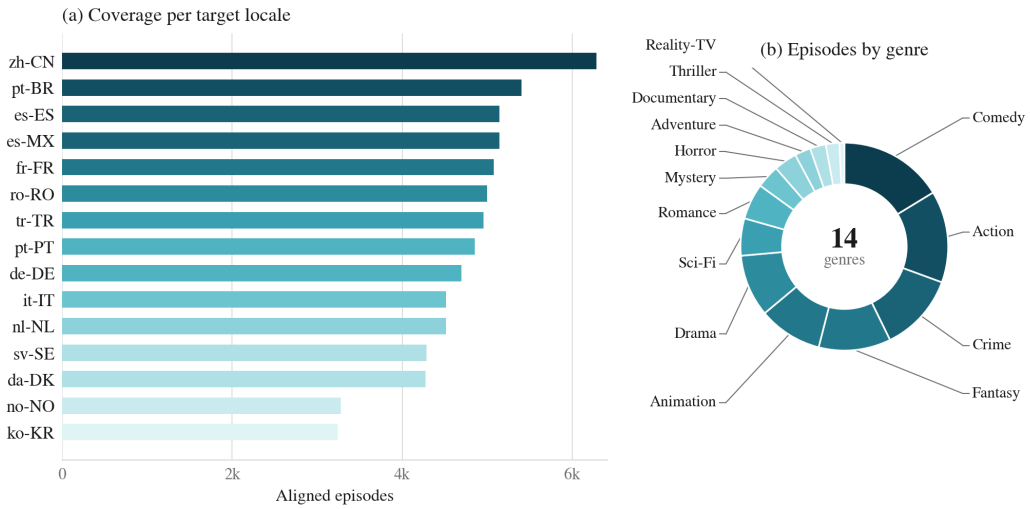


Figure 2: Visualization of **Subtitle Arena** statistics. **Left:** aligned bilingual episode coverage for each target locale, measured against 6,267 English source episodes. **Right:** distribution of the 192 television series across 14 genres.

D.6 Subtitle Arena Visualization

Figure 2 provides a visualization of the language coverage and genre composition summarized in Table 1.

E MQM Evaluation Protocol

We propose SubMQM by adapting Multidimensional Quality Metrics (MQM) to subtitle translation. Given a source subtitle and its translation, an LLM judge evaluates a sliding window of consecutive sentences, for each of the 19 subcategories in Table 10, a discrete error penalty in $\{0, 5, 10\}$: 0 indicates no error in that subcategory within the window, 5 a minor issue that does not impede comprehension, and 10 a severe error that changes meaning, misleads the viewer, or violates a hard subtitle constraint. Subcategories that are not applicable to a window (e.g., no profanity, no units) default to 0. All scores are penalties, so lower is better.

Aggregation. A dimension penalty is the mean of its subcategory penalties, averaged over all windows of an episode. The Overall penalty is a weighted average of the seven dimensions,

$$\text{Overall} = \textstyle\sum_d w_d\, s_d, \quad \sum_d w_d = 1,$$

with weights that emphasize semantic fidelity over surface form: Accuracy 0.3, Terminology 0.2, Fluency 0.2, Audience Appropriateness 0.12, Linguistic Conventions 0.08, Technical 0.06, and Locale Conventions 0.04. A lower Overall indicates fewer and less severe errors.

Alignment. Model outputs preserve the source block segmentation and are scored block-by-block. Human reference subtitles are segmented independently of the source, so we align them to the source by timestamp overlap and score at the passage level, preventing benign re-segmentation from being penalized as omission or mistranslation.

F Additional Experimental Results

F.1 English-to-Locale Results

Tables 11 and 12 provide the complete fine-grained English→locale results corresponding to the representative summary in Table 3. Across the 15 target locales, SMART achieves the lowest Overall SubMQM penalty in every direction. Its mean Overall penalty is 1.11, compared with 1.20 for TransAgent, corresponding to a 6.9% relative reduction. The reduction is not confined to one

Table 10: Subtitle-adapted MQM rubric used by SubMQM. The rubric contains seven dimensions and 19 error types. Each error type receives a penalty in $\{0, 5, 10\}$; lower is better.

Dimension	Error Type	Description
Terminology	Name Inconsistency	Inconsistent translations or spellings of recurring proper names, such as characters, places, and entities.
	Term Inconsistency	Inconsistent translations of recurring domain-specific or contextual terms and phrases.
Accuracy	Mistranslation	Translation that incorrectly changes the meaning of the source.
	Undertranslation	Source content that should be translated is omitted.
	Overtranslation	Unsupported information or specificity is introduced into the translation.
Fluency	Coherence	The translation lacks continuity with the surrounding scene or discourse.
	Naturalness	Awkward or translation-like phrasing that does not resemble native usage.
	Vividness	Emotional nuance, humor, wordplay, or stylistic expressiveness is weakened.
Linguistic Conventions	Mispunctuation	Missing, incorrect, or improperly formatted punctuation.
	Miscapitalization	Incorrect capitalization of sentence-initial words, proper nouns, or other forms.
	Grammar	Grammatical errors in agreement, morphology, syntax, or related constructions.
	Spacing Error	Missing, extra, or duplicated spaces around words or punctuation.
Technical	Incorrect Line Breaking	Line breaks improperly separate tightly coupled linguistic units.
	Exceeding Characters per Line	A subtitle line exceeds the predefined character-per-line constraint.
	Exceeding Lines per Box	A subtitle event exceeds the predefined maximum number of lines.
Locale Conventions	Localization Error	Units, currencies, dates, or culturally specific expressions are improperly localized.
	Language Detection Error	The detected language does not match the expected source or target language.
Audience Appropriateness	Profanity	Profanity is unjustifiably strengthened, weakened, introduced, or removed.
	Formality Error	The register or level of formality is inappropriate or inconsistent with the scene.

category: the mean Accuracy penalty decreases from 1.83 to 1.67, the mean Fluency penalty from 1.88 to 1.79, and the mean Technical penalty from 1.83 to 1.72.

The fine-grained breakdown further shows reductions in recurring long-form errors such as mistranslation, undertranslation, overtranslation, terminology inconsistency, and line-formatting violations. These improvements are consistent with the components targeted by SMART: persistent series memory maintains cross-episode information, contextual retrieval supplies relevant long-range evidence, and the judge–refiner loop revisits translations when local hypotheses conflict with accumulated context or subtitle constraints.

Table 11: Fine-grained SubMQM results for translation from English into 15 target locales: Terminology, Accuracy, and Fluency. All values are penalties (lower is better). Best and second-best values within each language direction and metric are highlighted with blue-gray and warm-beige backgrounds, respectively. These results provide the full breakdown corresponding to Table 3.

Dir.	Method	Terminology		Accuracy			Fluency			Overall
		NameInc	TermInc	MisTrans	UndTrans	OvrTrans	Coher	Natur	Vivid	
	Online	0.40	0.30	4.40	5.60	4.60	6.30	5.00	0.90	2.58
	Gemma 3 4B	0.52	0.46	3.45	2.68	2.15	2.65	2.12	1.15	1.46
	DeepSeek-V3.2	0.31	0.25	2.52	1.76	1.30	1.88	1.29	**0.58**	0.93
	Claude 4.6	0.26	0.25	2.31	1.43	1.15	2.03	**0.90**	**0.56**	0.84
en→zh	Claude 4.8	0.27	0.22	2.13	1.31	1.23	1.40	1.08	0.78	0.78

Continued on next page

Table 11 (continued)

Dir.	Method	Terminology		Accuracy			Fluency			Overall
		NameInc	TermInc	MisTrans	UndTrans	OvrTrans	Coher	Natur	Vivid	
	GPT-5.5	**0.21**	0.18	1.50	1.03	0.94	1.53	1.15	0.72	0.66
	DRT	0.26	0.18	1.41	0.95	0.96	1.33	**0.84**	0.94	0.61
	TransAgent	**0.18**	**0.13**	**0.92**	**0.71**	**0.92**	**1.05**	1.21	0.91	**0.52**
	SMART	**0.18**	**0.12**	**0.86**	**0.81**	**0.73**	**1.08**	0.94	0.83	**0.48**
	Online	0.64	0.38	8.35	6.52	5.20	3.48	2.91	1.89	2.77
	Gemma 3 4B	0.78	0.62	4.85	4.12	3.56	3.45	2.89	1.88	2.21
	DeepSeek-V3.2	0.50	0.37	3.52	2.81	2.38	2.54	1.98	1.25	1.50
	Claude 4.6	0.51	0.31	2.92	2.65	2.12	2.56	1.47	1.40	1.36
en→de	Claude 4.8	0.47	0.32	2.49	2.21	2.22	1.73	1.74	**1.16**	1.25
	GPT-5.5	0.40	0.31	2.42	1.86	1.93	2.40	1.62	**1.12**	1.21
	DRT	**0.28**	**0.19**	1.71	1.69	1.47	**1.72**	1.89	1.26	1.02
	TransAgent	0.31	**0.20**	**1.62**	**1.51**	**1.41**	1.79	**1.38**	1.22	**0.95**
	SMART	**0.29**	0.21	**1.39**	**1.28**	**1.05**	**1.62**	**1.46**	1.30	**0.87**
	Online	0.54	0.28	3.92	2.65	2.16	2.38	1.91	1.32	1.48
	Gemma 3 4B	0.75	0.61	3.82	3.12	2.74	2.98	2.54	1.78	1.82
	DeepSeek-V3.2	0.48	0.35	2.58	1.98	1.72	2.08	1.66	**1.14**	1.17
	Claude 4.6	0.47	0.33	2.51	1.89	1.70	**1.25**	**1.57**	1.15	1.07
en→ko	Claude 4.8	0.37	0.32	1.87	2.01	1.53	1.91	1.64	**1.12**	1.05
	GPT-5.5	0.44	0.30	2.45	1.85	**1.26**	1.48	1.82	**1.12**	1.04
	DRT	**0.34**	**0.21**	2.30	**1.35**	1.48	**1.44**	1.61	1.46	0.99
	TransAgent	**0.35**	0.30	**1.66**	1.78	1.44	1.58	1.68	1.44	**0.97**
	SMART	**0.34**	**0.24**	**1.76**	**1.56**	**1.33**	1.56	**1.41**	1.32	**0.92**
	Online	0.68	0.40	7.95	6.08	5.05	3.79	3.25	2.08	2.85
	Gemma 3 4B	0.78	0.62	4.95	4.15	3.65	3.98	3.42	2.35	2.34
	DeepSeek-V3.2	0.50	0.36	3.60	2.79	2.44	3.08	2.59	**1.71**	1.62
	Claude 4.6	0.41	0.34	3.20	2.21	2.21	2.90	2.66	**1.71**	1.49
en→it	Claude 4.8	0.45	0.28	2.45	2.34	2.05	2.89	2.09	2.31	1.39
	GPT-5.5	0.39	0.24	2.70	1.84	1.85	3.30	**2.02**	1.98	1.37
	DRT	0.35	0.19	**1.82**	**1.60**	**1.55**	2.58	**1.99**	2.13	1.19
	TransAgent	**0.27**	**0.15**	2.00	1.63	1.57	**2.54**	2.24	**1.32**	**1.17**
	SMART	**0.26**	**0.16**	**1.66**	**1.46**	**1.35**	**2.36**	2.12	1.97	**1.13**
	Online	0.66	0.40	8.38	6.65	5.40	3.91	3.15	2.03	2.90
	Gemma 3 4B	0.82	0.65	5.12	4.28	3.76	3.82	3.28	2.25	2.36
	DeepSeek-V3.2	0.52	0.38	3.70	2.89	2.51	2.86	2.38	1.60	1.62
	Claude 4.6	0.50	0.37	2.92	3.03	2.21	3.08	2.27	1.87	1.58
en→es-ES	Claude 4.8	0.50	**0.35**	3.48	2.55	2.49	2.88	2.09	1.58	1.55
	GPT-5.5	0.56	0.49	2.75	2.33	2.39	2.47	2.14	**1.48**	1.44
	DRT	0.49	**0.35**	**2.48**	2.46	1.96	**1.91**	2.09	1.54	1.34
	TransAgent	**0.40**	**0.34**	2.72	**1.69**	**1.61**	**1.51**	**2.02**	**1.26**	**1.21**
	SMART	**0.44**	**0.34**	**1.96**	**1.76**	**1.59**	2.06	**1.80**	1.66	**1.18**
	Online	0.46	0.24	7.48	5.65	4.66	3.89	3.05	2.15	2.61
	Gemma 3 4B	0.72	0.56	4.90	4.08	3.58	3.95	3.38	2.32	2.31
	DeepSeek-V3.2	0.42	0.28	3.55	2.80	2.40	3.04	2.50	1.69	1.60
	Claude 4.6	0.37	0.24	3.17	2.87	2.44	2.52	2.51	**1.61**	1.50
en→fr	Claude 4.8	0.32	0.22	3.04	2.63	2.10	2.51	**1.87**	1.75	1.44
	GPT-5.5	0.32	0.16	2.49	2.57	1.95	2.96	2.38	1.97	1.43
	DRT	0.25	**0.14**	2.21	**1.85**	1.95	2.89	2.51	**1.68**	1.31
	TransAgent	**0.22**	**0.10**	**1.86**	1.88	**1.14**	**2.36**	2.20	1.97	**1.20**
	SMART	**0.20**	**0.10**	**1.82**	**1.58**	**1.43**	**2.41**	**2.13**	1.97	**1.19**
	Online	1.57	1.18	8.38	6.56	6.11	7.26	6.02	6.48	4.51
	Gemma 3 4B	1.38	1.04	6.80	5.06	5.47	6.44	4.75	5.14	3.72
	DeepSeek-V3.2	1.01	0.84	5.05	3.62	3.70	4.73	3.70	4.07	2.73
	Claude 4.6	0.88	0.69	4.15	3.10	3.03	3.68	3.38	3.32	2.29
en→pt-PT	Claude 4.8	0.79	0.61	3.66	2.86	2.65	3.49	2.98	3.00	2.05
	GPT-5.5	0.72	0.54	3.33	2.58	2.38	3.23	2.83	2.65	1.86
	DRT	0.65	0.48	2.92	2.39	2.08	2.88	2.69	2.36	1.67
	TransAgent	**0.58**	**0.41**	**2.55**	**2.11**	**1.97**	**2.64**	**2.47**	**2.15**	**1.50**
	SMART	**0.53**	**0.37**	**2.41**	**1.93**	**1.83**	**2.43**	**2.18**	**2.03**	**1.39**
	Online	1.34	0.91	4.80	4.47	3.65	5.02	4.19	4.00	2.97
	Gemma 3 4B	1.04	0.78	4.24	3.72	3.26	4.30	3.69	3.60	2.56
	DeepSeek-V3.2	0.71	0.53	3.27	2.88	2.52	3.29	2.73	2.52	1.91
	Claude 4.6	0.58	0.50	2.76	2.33	2.02	3.11	2.52	2.17	1.64
en→pt-BR	Claude 4.8	0.51	0.44	2.61	2.07	1.77	2.71	2.27	2.04	1.48
	GPT-5.5	0.49	0.41	2.36	1.85	1.57	2.39	2.14	1.86	1.34
	DRT	0.44	0.38	2.08	1.72	1.47	2.23	2.00	1.73	1.23
	TransAgent	**0.41**	**0.35**	**1.93**	**1.59**	**1.39**	**2.01**	**1.91**	**1.54**	**1.13**
	SMART	**0.36**	**0.32**	**1.79**	**1.48**	**1.26**	**1.85**	**1.69**	**1.45**	**1.04**
	Online	1.59	1.19	5.03	5.99	4.88	6.44	6.93	4.34	3.60
	Gemma 3 4B	1.30	0.97	4.32	4.99	3.96	5.22	5.42	3.57	2.99
	DeepSeek-V3.2	0.93	0.76	3.33	3.37	3.04	3.96	3.64	2.54	2.18
	Claude 4.6	0.74	0.59	2.87	2.91	2.72	3.35	3.16	2.38	1.89
en→no	Claude 4.8	0.66	0.53	2.65	2.64	2.53	2.94	2.82	2.21	1.71
	GPT-5.5	0.61	0.46	2.38	2.47	2.27	2.64	2.67	2.02	1.56
	DRT	0.57	0.42	2.20	2.19	2.17	2.38	2.43	1.83	1.43

Continued on next page

Table 11 (continued)

Dir.	Method	Terminology		Accuracy			Fluency			Overall
		NameInc	TermInc	MisTrans	UndTrans	OvrTrans	Coher	Natur	Vivid	
	TransAgent	**0.53**	**0.37**	**2.04**	**2.04**	**1.91**	**2.17**	**2.24**	**1.66**	**1.31**
	SMART	**0.46**	**0.33**	**1.78**	**1.86**	**1.74**	**2.05**	**1.99**	**1.53**	**1.19**
en→da	Online	1.34	1.00	7.79	5.42	4.51	7.83	6.58	5.36	3.99
	Gemma 3 4B	1.10	0.88	6.12	4.72	3.83	6.45	5.29	4.23	3.30
	DeepSeek-V3.2	0.84	0.60	4.31	3.22	3.00	4.70	3.56	2.93	2.34
	Claude 4.6	0.74	0.55	3.36	3.00	2.46	3.73	2.79	2.69	1.97
	Claude 4.8	0.70	0.49	3.13	2.82	2.26	3.54	2.62	2.50	1.83
	GPT-5.5	0.66	0.45	2.82	2.57	2.07	3.11	2.41	2.22	1.66
	DRT	0.58	0.41	2.55	2.42	1.95	2.70	2.11	1.93	1.50
	TransAgent	**0.54**	**0.36**	**2.43**	**2.16**	**1.85**	**2.38**	**1.97**	**1.76**	**1.38**
	SMART	**0.50**	**0.34**	**2.30**	**1.88**	**1.74**	**2.12**	**1.80**	**1.61**	**1.26**
en→nl	Online	1.53	1.26	6.40	5.09	6.66	6.20	6.38	6.56	4.11
	Gemma 3 4B	1.21	1.08	5.75	4.02	5.25	5.54	5.17	5.14	3.41
	DeepSeek-V3.2	0.82	0.77	4.67	3.17	3.74	4.17	3.99	3.71	2.57
	Claude 4.6	0.77	0.61	3.80	2.95	3.13	3.67	3.33	2.94	2.17
	Claude 4.8	0.68	0.56	3.44	2.66	2.95	3.33	3.03	2.72	1.99
	GPT-5.5	0.65	0.50	3.06	2.45	2.57	3.14	2.84	2.60	1.81
	DRT	0.61	0.46	2.85	2.27	2.42	2.96	2.60	2.45	1.68
	TransAgent	**0.55**	**0.44**	**2.63**	**2.06**	**2.16**	**2.61**	**2.30**	**2.25**	**1.52**
	SMART	**0.52**	**0.40**	**2.36**	**1.91**	**1.88**	**2.44**	**2.14**	**2.02**	**1.38**
en→ro	Online	1.29	1.11	6.32	4.69	4.71	6.61	4.53	5.31	3.48
	Gemma 3 4B	1.12	0.94	5.01	4.20	4.23	5.61	3.79	4.67	2.96
	DeepSeek-V3.2	0.81	0.65	3.50	3.13	3.24	3.80	2.95	3.51	2.17
	Claude 4.6	0.67	0.52	2.95	2.43	2.61	3.00	2.40	2.75	1.75
	Claude 4.8	0.59	0.45	2.67	2.20	2.35	2.70	2.25	2.39	1.59
	GPT-5.5	0.53	0.42	2.37	2.07	2.05	2.50	1.98	2.10	1.42
	DRT	0.50	0.37	2.16	1.91	1.89	2.22	1.77	1.88	1.29
	TransAgent	**0.46**	**0.35**	**2.03**	**1.67**	**1.70**	**1.96**	**1.54**	**1.67**	**1.17**
	SMART	**0.43**	**0.33**	**1.84**	**1.58**	**1.56**	**1.84**	**1.46**	**1.51**	**1.07**
en→sv	Online	1.16	0.95	4.94	4.97	4.53	5.12	3.83	4.94	3.15
	Gemma 3 4B	0.93	0.78	3.86	4.33	3.62	3.99	3.16	4.01	2.57
	DeepSeek-V3.2	0.77	0.56	3.14	2.91	2.59	3.31	2.44	2.75	1.91
	Claude 4.6	0.61	0.44	2.97	2.48	2.22	2.88	2.22	2.22	1.65
	Claude 4.8	0.57	0.41	2.60	2.34	2.08	2.51	2.09	1.95	1.51
	GPT-5.5	0.50	0.38	2.46	2.19	1.90	2.36	1.91	1.76	1.39
	DRT	0.45	0.33	2.19	1.96	1.72	2.24	1.80	1.65	1.27
	TransAgent	**0.40**	**0.30**	**2.05**	**1.79**	**1.53**	**1.96**	**1.70**	**1.50**	**1.15**
	SMART	**0.36**	**0.28**	**1.81**	**1.60**	**1.34**	**1.76**	**1.61**	**1.36**	**1.04**
en→tr	Online	1.34	1.10	7.30	4.85	4.01	6.14	5.78	4.84	3.60
	Gemma 3 4B	1.18	0.85	5.78	4.39	3.64	5.10	4.48	4.30	3.04
	DeepSeek-V3.2	0.93	0.62	4.18	3.14	2.58	3.94	3.46	3.45	2.26
	Claude 4.6	0.74	0.59	3.39	2.97	2.30	3.18	3.01	2.93	1.94
	Claude 4.8	0.64	0.56	3.11	2.76	2.06	3.01	2.61	2.74	1.77
	GPT-5.5	0.60	0.50	2.78	2.56	1.93	2.66	2.46	2.44	1.61
	DRT	0.54	0.45	2.43	2.41	1.82	2.41	2.15	2.27	1.47
	TransAgent	**0.48**	**0.40**	**2.31**	**2.12**	**1.66**	**2.23**	**1.99**	**1.99**	**1.33**
	SMART	**0.45**	**0.38**	**2.19**	**1.98**	**1.55**	**2.01**	**1.83**	**1.80**	**1.25**
en→es-MX	Online	1.49	1.13	6.32	8.42	5.82	6.23	5.06	5.69	4.11
	Gemma 3 4B	1.17	1.00	5.30	6.87	4.85	5.40	4.58	4.67	3.43
	DeepSeek-V3.2	0.94	0.69	4.05	4.73	3.28	3.72	3.39	3.86	2.51
	Claude 4.6	0.85	0.63	3.25	3.74	2.95	3.04	3.07	3.34	2.12
	Claude 4.8	0.76	0.58	3.07	3.28	2.57	2.69	2.91	2.91	1.91
	GPT-5.5	0.68	0.54	2.80	3.07	2.37	2.49	2.62	2.60	1.75
	DRT	0.59	0.48	2.45	2.71	2.08	2.32	2.31	2.27	1.56
	TransAgent	**0.55**	**0.42**	**2.28**	**2.35**	**1.97**	**2.21**	**2.15**	**2.08**	**1.43**
	SMART	**0.50**	**0.38**	**2.15**	**2.05**	**1.73**	**2.08**	**2.01**	**1.97**	**1.31**

Abbreviations. Terminology (NameInc = Name Inconsistency; TermInc = Term Inconsistency). **Accuracy** (MisTrans = Mistranslation; UndTrans = Undertranslation; OvrTrans = Overtranslation). **Fluency** (Coher = Coherence; Natur = Naturalness; Vivid = Vividness).

Table 12: Fine-grained SubMQM results for translation from English into 15 target locales: Linguistic Conventions, Technical, Locale Conventions, and Audience Appropriateness. All values are penalties (lower is better). Best and second-best values within each language direction and metric are highlighted with blue-gray and warm-beige backgrounds, respectively. These results provide the full breakdown corresponding to Table 3.

Dir.	Method	Linguistic Conventions				Technical			Locale Conventions		Audience Appropriateness		Overall
		MisPunc	MisCap	Gram	Space	LineBrk	CharLim	LineLim	LocErr	LangDet	Profan	Formal	
	Online	4.30	**0.00**	0.50	3.10	0.10	0.50	**0.00**	**0.00**	2.80	**0.10**	**0.00**	2.58
	Gemma 3 4B	1.05	**0.00**	0.58	0.72	1.18	0.84	0.35	0.18	0.12	0.38	0.31	1.46
	DeepSeek-V3.2	0.61	**0.00**	0.19	0.34	0.82	0.52	0.14	**0.04**	**0.01**	0.15	0.10	0.93
en→zh													

Continued on next page

Table 12 (continued)

Dir.	Method	Linguistic Conventions				Technical			Locale Conventions		Audience Appropriateness		Overall
		MisPunc	MisCap	Gram	Space	LineBrk	CharLim	LineLim	LocErr	LangDet	Profan	Formal	
en→de	Claude 4.6	0.55	**0.00**	0.16	0.31	0.70	0.44	0.12	**0.04**	**0.02**	0.20	0.11	0.84
	Claude 4.8	0.41	**0.00**	0.13	0.26	0.53	0.34	0.09	0.05	**0.02**	0.15	0.10	0.78
	GPT-5.5	0.37	**0.00**	0.12	0.27	0.37	0.23	0.06	**0.04**	**0.02**	0.18	0.10	0.66
	DRT	0.24	**0.00**	0.08	**0.16**	0.20	0.13	0.04	0.06	**0.02**	**0.11**	0.10	0.61
	TransAgent	**0.15**	**0.00**	**0.06**	**0.13**	**0.08**	**0.05**	**0.01**	**0.04**	0.03	0.19	**0.08**	**0.52**
	SMART	**0.10**	**0.00**	**0.05**	**0.13**	**0.00**	**0.00**	**0.00**	0.05	0.03	0.16	**0.08**	**0.48**
	Online	1.22	0.39	0.54	0.29	**0.08**	**0.05**	**0.02**	1.31	0.51	**0.23**	**0.09**	2.77
	Gemma 3 4B	1.24	0.65	0.86	0.68	2.45	1.98	0.88	0.68	0.28	0.64	0.52	2.21
	DeepSeek-V3.2	0.70	0.27	0.41	0.26	2.02	1.61	0.52	0.38	0.05	0.36	0.25	1.50
	Claude 4.6	0.75	0.25	0.39	0.25	**1.48**	1.72	**0.34**	0.33	0.05	0.33	0.29	1.36
	Claude 4.8	0.56	0.24	0.43	**0.23**	2.19	1.74	0.83	0.26	0.04	**0.32**	**0.24**	1.25
	GPT-5.5	0.47	0.23	0.35	0.24	1.87	1.74	1.14	0.19	**0.03**	0.44	0.31	1.21
	DRT	0.39	0.17	**0.31**	**0.23**	2.08	**1.41**	1.27	0.12	**0.03**	0.43	0.26	1.02
	TransAgent	**0.33**	**0.15**	0.34	0.24	1.77	1.50	0.65	**0.07**	**0.02**	0.44	0.35	**0.95**
	SMART	**0.28**	**0.14**	**0.27**	**0.19**	1.85	1.64	1.16	**0.04**	**0.02**	0.48	0.34	**0.87**
en→ko	Online	1.28	**0.00**	0.74	0.50	0.41	0.29	0.11	**0.06**	**0.02**	0.89	**0.51**	1.48
	Gemma 3 4B	1.02	**0.00**	0.82	0.71	1.14	0.82	0.45	0.28	0.15	1.15	0.92	1.82
	DeepSeek-V3.2	0.55	**0.00**	0.38	0.30	0.74	0.49	0.18	**0.06**	**0.02**	0.76	**0.56**	1.17
	Claude 4.6	0.49	**0.00**	0.38	0.27	0.71	0.44	0.16	**0.06**	**0.02**	**0.66**	**0.56**	1.07
	Claude 4.8	0.49	**0.00**	0.34	0.26	0.55	0.35	0.14	**0.05**	**0.03**	0.78	0.65	1.05
	GPT-5.5	0.34	**0.00**	0.33	**0.20**	0.40	0.26	0.10	**0.05**	**0.02**	0.79	0.57	1.04
	DRT	0.30	**0.00**	**0.28**	0.28	0.32	0.18	0.09	**0.06**	**0.02**	0.80	0.69	0.99
	TransAgent	**0.25**	**0.00**	**0.28**	**0.21**	**0.22**	**0.11**	**0.06**	**0.06**	**0.03**	**0.73**	0.60	**0.97**
	SMART	**0.26**	**0.00**	**0.22**	**0.20**	**0.17**	**0.08**	**0.05**	**0.05**	**0.03**	0.93	0.65	**0.92**
en→it	Online	3.48	1.72	2.35	1.29	**0.24**	**0.13**	**0.02**	0.96	0.40	**0.10**	**0.04**	2.85
	Gemma 3 4B	1.38	0.78	1.05	0.82	2.38	1.88	0.82	0.62	0.25	0.68	0.56	2.34
	DeepSeek-V3.2	0.86	0.39	0.58	0.36	1.91	**1.48**	**0.48**	0.34	0.06	0.34	0.23	1.62
	Claude 4.6	0.89	0.34	0.54	0.34	1.92	1.68	0.85	0.35	**0.05**	0.32	0.22	1.49
	Claude 4.8	0.67	0.31	0.45	0.29	**1.70**	1.67	0.71	0.33	**0.04**	**0.30**	0.23	1.39
	GPT-5.5	0.60	0.21	0.41	0.34	2.43	1.56	1.46	0.23	0.06	0.41	**0.21**	1.37
	DRT	0.51	0.21	0.44	0.32	2.51	2.12	1.04	0.20	0.06	0.42	0.25	1.19
	TransAgent	**0.38**	**0.18**	**0.34**	**0.24**	2.85	2.56	1.44	**0.19**	0.07	0.45	**0.21**	**1.17**
	SMART	**0.37**	**0.17**	**0.33**	**0.25**	2.76	2.44	1.91	**0.16**	0.06	0.43	0.27	**1.13**
en→es-ES	Online	2.28	1.12	1.39	0.77	**0.06**	**0.02**	**0.01**	0.76	0.26	**0.17**	**0.05**	2.90
	Gemma 3 4B	1.42	0.88	1.12	0.86	2.32	1.82	0.78	0.58	0.22	0.71	0.58	2.36
	DeepSeek-V3.2	0.91	0.49	0.64	0.41	1.85	**1.42**	**0.44**	0.26	**0.05**	0.36	**0.25**	1.62
	Claude 4.6	0.91	0.45	0.64	**0.32**	1.92	2.02	0.89	0.22	**0.05**	0.40	0.31	1.58
	Claude 4.8	1.01	0.42	0.54	**0.33**	2.35	1.67	0.63	0.21	**0.05**	**0.29**	0.32	1.55
	GPT-5.5	**0.63**	0.40	**0.47**	0.47	**1.83**	1.81	0.93	0.19	**0.05**	0.59	**0.25**	1.44
	DRT	0.68	0.36	0.50	0.38	2.21	2.13	1.36	0.14	**0.04**	0.38	0.29	1.34
	TransAgent	**0.60**	**0.32**	**0.37**	0.41	2.54	2.39	1.18	**0.11**	**0.05**	0.49	0.30	**1.21**
	SMART	0.66	**0.26**	0.52	0.40	2.31	1.99	1.49	**0.10**	**0.04**	0.51	0.35	**1.18**
en→fr	Online	2.56	1.21	1.83	0.88	**0.12**	**0.06**	**0.03**	0.28	0.06	**0.21**	**0.09**	2.61
	Gemma 3 4B	1.40	0.82	1.08	0.84	2.48	1.95	0.85	0.48	0.24	0.74	0.60	2.31
	DeepSeek-V3.2	0.90	0.40	0.60	0.35	2.01	**1.55**	0.52	0.16	**0.03**	0.40	0.27	1.60
	Claude 4.6	0.90	0.36	0.60	0.34	**1.17**	1.73	**0.36**	0.16	**0.03**	**0.37**	0.28	1.50
	Claude 4.8	0.76	0.32	0.44	0.39	2.84	1.61	1.32	0.14	**0.03**	**0.37**	0.38	1.44
	GPT-5.5	0.67	0.38	0.48	0.33	1.47	2.59	1.21	0.12	**0.02**	0.46	0.30	1.43
	DRT	0.71	0.27	**0.43**	0.40	2.45	1.64	2.03	**0.10**	**0.03**	0.44	**0.26**	1.31
	TransAgent	**0.47**	**0.17**	0.51	**0.26**	3.37	2.35	2.26	**0.08**	**0.02**	0.59	0.34	**1.20**
	SMART	**0.52**	**0.18**	**0.42**	**0.32**	3.01	2.65	2.08	**0.08**	**0.02**	0.53	0.35	**1.19**
en→pt-PT	Online	3.08	1.05	1.47	1.61	8.83	8.89	5.73	0.45	0.17	1.68	1.39	4.51
	Gemma 3 4B	2.36	0.91	1.14	1.43	7.11	6.92	4.75	0.38	0.14	1.32	1.07	3.72
	DeepSeek-V3.2	1.66	0.68	0.91	0.96	5.43	5.24	3.38	0.25	0.11	1.01	0.82	2.73
	Claude 4.6	1.42	0.57	0.85	0.79	4.26	4.63	2.61	0.23	0.10	0.95	0.66	2.29
	Claude 4.8	1.25	0.51	0.80	0.71	3.84	4.07	2.32	0.20	0.09	0.89	0.59	2.05
	GPT-5.5	1.16	0.47	0.74	0.63	3.59	3.55	2.07	0.18	0.07	0.80	0.55	1.86
	DRT	1.04	0.40	0.66	0.59	3.13	3.12	1.89	0.16	0.06	0.72	0.50	1.67
	TransAgent	**0.90**	**0.36**	**0.62**	**0.52**	**2.80**	**2.72**	**1.72**	**0.14**	**0.05**	**0.67**	**0.47**	**1.50**
	SMART	**0.82**	**0.33**	**0.58**	**0.46**	**2.60**	**2.47**	**1.61**	**0.13**	**0.04**	**0.64**	**0.41**	**1.39**
en→pt-BR	Online	2.05	0.68	1.57	0.98	5.39	6.92	4.14	0.26	0.12	1.16	0.90	2.97
	Gemma 3 4B	1.86	0.52	1.26	0.83	4.47	5.61	3.70	0.21	0.11	0.98	0.80	2.56
	DeepSeek-V3.2	1.25	0.41	0.90	0.68	3.25	4.05	2.71	0.15	0.10	0.81	0.56	1.91
	Claude 4.6	1.02	0.36	0.78	0.61	3.01	3.16	2.11	0.14	0.09	0.71	0.50	1.64
	Claude 4.8	0.90	0.32	0.71	0.57	2.70	2.91	1.85	0.13	0.07	0.65	0.47	1.48
	GPT-5.5	0.78	0.30	0.63	0.51	2.46	2.76	1.66	0.12	0.06	0.62	0.42	1.34
	DRT	0.71	0.26	0.59	0.45	2.33	2.50	1.50	0.11	0.05	0.54	0.39	1.23
	TransAgent	**0.63**	**0.25**	**0.54**	**0.42**	**2.11**	**2.18**	**1.41**	**0.10**	**0.04**	**0.50**	**0.35**	**1.13**
	SMART	**0.60**	**0.23**	**0.50**	**0.37**	**2.00**	**1.90**	**1.26**	**0.09**	**0.03**	**0.45**	**0.31**	**1.04**
en→no	Online	2.13	0.76	1.79	1.16	5.58	4.86	3.66	0.30	0.13	1.39	1.02	3.60
	Gemma 3 4B	1.86	0.69	1.53	0.99	4.99	4.38	3.28	0.27	0.12	1.25	0.89	2.99
	DeepSeek-V3.2	1.45	0.52	1.11	0.70	3.61	3.49	2.32	0.21	0.11	0.84	0.68	2.18
	Claude 4.6	1.15	0.44	0.95	0.63	3.18	2.92	2.04	0.17	0.10	0.76	0.56	1.89
	Claude 4.8	1.04	0.40	0.85	0.59	2.94	2.69	1.90	0.15	0.08	0.68	0.53	1.71
	GPT-5.5	0.98	0.36	0.75	0.54	2.68	2.44	1.73	0.13	0.07	0.62	0.47	1.56
	DRT	0.88	0.32	0.69	0.52	2.55	2.25	1.65	0.12	0.06	0.56	0.44	1.43
	TransAgent	**0.80**	**0.29**	**0.63**	**0.45**	**2.35**	**2.15**	**1.53**	**0.11**	**0.05**	**0.50**	**0.39**	**1.31**
	SMART	**0.70**	**0.26**	**0.57**	**0.42**	**2.15**	**1.89**	**1.36**	**0.10**	**0.04**	**0.46**	**0.37**	**1.19**
en→da	Online	2.65	1.00	1.78	1.12	7.79	5.86	4.60	0.31	0.13	1.73	1.11	3.99
	Gemma 3 4B	2.29	0.77	1.47	0.97	6.96	4.70	4.01	0.25	0.12	1.41	0.93	3.30
	DeepSeek-V3.2	1.54	0.54	1.07	0.69	4.75	3.45	2.98	0.18	0.11	1.01	0.66	2.34

Continued on next page

Table 12 (continued)

Dir.	Method	Linguistic Conventions				Technical			Locale Conventions		Audience Appropriateness		Overall
		MisPunc	MisCap	Gram	Space	LineBrk	CharLim	LineLim	LocErr	LangDet	Profan	Formal	
	Claude 4.6	1.21	0.41	0.91	0.58	3.91	3.12	2.66	0.16	0.10	0.91	0.58	1.97
	Claude 4.8	1.08	0.39	0.82	0.55	3.52	2.75	2.46	0.14	0.08	0.82	0.54	1.83
	GPT-5.5	1.00	0.36	0.73	0.49	3.18	2.53	2.21	0.13	0.07	0.75	0.47	1.66
	DRT	0.89	0.34	0.68	0.43	2.86	2.36	1.92	0.12	0.06	0.67	0.43	1.50
	TransAgent	**0.84**	**0.31**	**0.59**	**0.41**	**2.53**	**2.16**	**1.80**	**0.11**	**0.05**	**0.58**	**0.41**	**1.38**
	SMART	**0.74**	**0.27**	**0.54**	**0.39**	**2.32**	**1.99**	**1.64**	**0.10**	**0.04**	**0.52**	**0.39**	**1.26**
	Online	2.31	0.84	1.78	1.47	6.59	7.80	5.23	0.37	0.13	2.03	1.49	4.11
	Gemma 3 4B	1.93	0.70	1.49	1.16	5.79	6.82	4.40	0.32	0.12	1.70	1.23	3.41
	DeepSeek-V3.2	1.40	0.52	1.04	0.88	4.71	5.00	3.22	0.21	0.11	1.14	0.88	2.57
	Claude 4.6	1.29	0.49	0.97	0.71	3.86	4.06	2.65	0.17	0.10	0.97	0.71	2.17
en→nl	Claude 4.8	1.18	0.43	0.85	0.67	3.65	3.76	2.33	0.15	0.08	0.85	0.67	1.99
	GPT-5.5	1.04	0.39	0.77	0.60	3.43	3.29	2.10	0.14	0.07	0.76	0.59	1.81
	DRT	0.92	0.36	0.68	0.53	3.14	2.98	1.86	0.13	0.06	0.68	0.52	1.68
	TransAgent	**0.85**	**0.33**	**0.61**	**0.50**	**2.75**	**2.64**	**1.76**	**0.12**	**0.05**	**0.60**	**0.47**	**1.52**
	SMART	**0.75**	**0.29**	**0.55**	**0.47**	**2.43**	**2.33**	**1.68**	**0.11**	**0.04**	**0.53**	**0.43**	**1.38**
	Online	1.95	0.81	1.93	1.16	7.46	4.56	4.16	0.28	0.12	1.31	0.95	3.48
	Gemma 3 4B	1.62	0.66	1.64	0.94	5.74	3.76	3.44	0.23	0.11	1.04	0.80	2.96
	DeepSeek-V3.2	1.17	0.53	1.16	0.63	4.43	3.06	2.48	0.18	0.10	0.81	0.60	2.17
	Claude 4.6	0.92	0.41	0.91	0.51	3.68	2.62	1.98	0.15	0.09	0.70	0.47	1.75
en→ro	Claude 4.8	0.84	0.38	0.84	0.47	3.49	2.38	1.75	0.13	0.07	0.65	0.44	1.59
	GPT-5.5	0.75	0.36	0.73	0.43	3.08	2.09	1.63	0.12	0.06	0.57	0.39	1.42
	DRT	0.67	0.31	0.66	0.40	2.73	1.93	1.52	0.11	0.05	0.54	0.36	1.29
	TransAgent	**0.62**	**0.28**	**0.57**	**0.38**	**2.46**	**1.77**	**1.40**	**0.10**	**0.04**	**0.51**	**0.31**	**1.17**
	SMART	**0.57**	**0.25**	**0.51**	**0.36**	**2.14**	**1.63**	**1.27**	**0.09**	**0.03**	**0.47**	**0.28**	**1.07**
	Online	1.63	0.51	1.31	1.05	7.52	5.40	4.10	0.24	0.12	1.13	1.05	3.15
	Gemma 3 4B	1.40	0.44	1.08	0.87	6.18	4.54	3.39	0.21	0.11	1.00	0.87	2.57
	DeepSeek-V3.2	0.95	0.37	0.79	0.63	4.22	3.53	2.50	0.17	0.10	0.77	0.58	1.91
	Claude 4.6	0.86	0.32	0.69	0.58	3.37	2.91	2.19	0.15	0.09	0.64	0.52	1.65
en→sv	Claude 4.8	0.76	0.30	0.63	0.53	2.97	2.59	2.07	0.14	0.07	0.59	0.48	1.51
	GPT-5.5	0.69	0.27	0.58	0.47	2.77	2.42	1.86	0.12	0.06	0.56	0.43	1.39
	DRT	0.63	0.24	0.53	0.42	2.49	2.27	1.66	0.11	0.05	0.51	0.39	1.27
	TransAgent	**0.56**	**0.22**	**0.48**	**0.38**	**2.27**	**1.98**	**1.56**	**0.10**	**0.04**	**0.44**	**0.35**	**1.15**
	SMART	**0.52**	**0.20**	**0.44**	**0.35**	**1.98**	**1.76**	**1.38**	**0.09**	**0.03**	**0.41**	**0.32**	**1.04**
	Online	1.89	0.73	1.50	1.14	6.66	7.49	4.35	0.31	0.13	1.57	0.92	3.60
	Gemma 3 4B	1.63	0.58	1.34	0.92	5.51	5.80	3.73	0.26	0.11	1.37	0.79	3.04
	DeepSeek-V3.2	1.15	0.41	0.94	0.61	4.16	3.96	2.97	0.21	0.10	1.01	0.63	2.26
	Claude 4.6	1.06	0.38	0.85	0.55	3.55	3.35	2.46	0.19	0.09	0.88	0.57	1.94
en→tr	Claude 4.8	0.95	0.35	0.79	0.52	3.12	3.04	2.21	0.18	0.08	0.77	0.54	1.77
	GPT-5.5	0.87	0.30	0.71	0.48	2.94	2.69	2.03	0.15	0.07	0.70	0.50	1.61
	DRT	0.82	0.28	0.63	0.45	2.73	2.47	1.82	0.13	0.06	0.64	0.46	1.47
	TransAgent	**0.77**	**0.26**	**0.56**	**0.40**	**2.45**	**2.15**	**1.60**	**0.12**	**0.05**	**0.59**	**0.43**	**1.33**
	SMART	**0.71**	**0.25**	**0.51**	**0.38**	**2.27**	**2.04**	**1.47**	**0.11**	**0.04**	**0.52**	**0.40**	**1.25**
	Online	1.95	0.92	1.73	1.37	8.42	6.82	4.75	0.36	0.17	1.27	0.90	4.11
	Gemma 3 4B	1.68	0.74	1.37	1.22	6.58	5.71	3.91	0.28	0.14	1.07	0.79	3.43
	DeepSeek-V3.2	1.31	0.50	1.03	0.87	5.19	4.27	2.92	0.20	0.11	0.88	0.64	2.51
	Claude 4.6	1.14	0.47	0.96	0.72	4.43	3.29	2.43	0.15	0.10	0.78	0.59	2.12
en→es-MX	Claude 4.8	1.00	0.43	0.87	0.65	4.09	2.87	2.16	0.13	0.09	0.71	0.54	1.91
	GPT-5.5	0.95	0.39	0.79	0.60	3.57	2.65	2.00	0.12	0.08	0.68	0.47	1.75
	DRT	0.87	0.36	0.72	0.54	3.17	2.46	1.83	0.11	0.07	0.63	0.41	1.56
	TransAgent	**0.82**	**0.32**	**0.68**	**0.49**	**2.81**	**2.33**	**1.75**	**0.10**	**0.06**	**0.58**	**0.39**	**1.43**
	SMART	**0.76**	**0.29**	**0.63**	**0.45**	**2.47**	**2.21**	**1.59**	**0.09**	**0.05**	**0.52**	**0.36**	**1.31**

Abbreviations. Linguistic Conventions (MisPunc = Mispunctuation; MisCap = Miscapitalization; Gram = Grammar; Space = Spacing Error). **Technical** (LineBrk = Incorrect Line Breaking; CharLim = Exceeding Characters per Line; LineLim = Exceeding Lines per Box). **Locale Conventions** (LocErr = Localization Error; LangDet = Language Detection Error). **Audience Appropriateness** (Profan = Profanity; Formal = Formality Error).

F.2 Locale-to-English Results

Tables 13 and 14 provide the complete locale→English results corresponding to Table 3. SMART again achieves the lowest Overall SubMQM penalty in all 15 directions, with a mean Overall penalty of 0.37 compared with 0.41 for TransAgent, a 10.0% relative reduction. The mean Accuracy penalty decreases from 0.75 to 0.67, while the mean Technical penalty decreases from 0.13 to 0.07; Terminology, Fluency, Linguistic Conventions, Locale Conventions, and Audience Appropriateness also improve on average.

The error-type breakdown is not uniformly better in every individual category: for example, Vividness and Profanity show small mixed changes in some directions. The Overall improvement instead comes from consistent reductions across the higher-weight semantic and consistency errors, particularly mistranslation, undertranslation, overtranslation, coherence, terminology inconsistency, and subtitle-formatting violations. Because every system generates English in this setting, these results also show that SMART's gains are not specific to target-language morphology or typography.

Table 13: Fine-grained SubMQM results for translation from 15 source locales into English: Terminology, Accuracy, and Fluency. All values are penalties (lower is better). Best and second-best values within each language direction and metric are highlighted with blue-gray and warm-beige backgrounds, respectively. These results provide the full breakdown corresponding to Table 3.

Dir.	Method	Terminology		Accuracy			Fluency			Overall
		NameInc	TermInc	MisTrans	UndTrans	OvrTrans	Coher	Natur	Vivid	
zh→en	Online	0.47	0.36	4.92	4.05	3.71	4.56	3.79	1.49	2.17
	Gemma 3 4B	0.58	0.45	3.91	2.89	2.51	2.91	2.24	1.34	1.64
	DeepSeek-V3.2	0.36	0.26	2.74	1.86	1.50	2.11	1.39	0.78	1.07
	Claude 4.6	0.32	0.22	2.49	1.68	1.34	1.89	1.26	0.69	0.96
	Claude 4.8	0.24	0.16	1.86	1.24	1.08	1.49	0.98	0.54	0.73
	GPT-5.5	0.21	0.14	1.59	1.06	0.94	1.32	1.04	0.58	0.67
	DRT	0.19	0.13	1.28	0.91	0.84	1.19	0.88	**0.52**	0.55
	TransAgent	**0.17**	**0.11**	**1.16**	**0.82**	**0.77**	**1.10**	**0.80**	**0.46**	**0.50**
	SMART	**0.15**	**0.10**	**0.95**	**0.74**	**0.71**	**1.02**	**0.78**	0.56	**0.44**
de→en	Online	0.44	0.33	4.31	3.59	3.16	3.26	2.69	1.39	1.80
	Gemma 3 4B	0.53	0.39	3.58	2.68	2.32	2.60	1.99	1.18	1.49
	DeepSeek-V3.2	0.35	0.22	2.50	1.72	1.41	1.88	1.26	0.71	1.02
	Claude 4.6	0.30	0.20	2.29	1.58	1.26	1.69	1.16	0.62	0.88
	Claude 4.8	0.22	0.14	1.72	1.14	1.01	1.36	0.91	0.49	0.67
	GPT-5.5	0.19	0.13	1.46	0.98	0.88	1.22	0.94	0.52	0.62
	DRT	0.18	**0.11**	1.19	0.84	0.78	1.09	0.80	**0.46**	0.50
	TransAgent	**0.15**	**0.11**	**1.08**	**0.75**	**0.71**	**1.00**	**0.72**	**0.42**	**0.45**
	SMART	**0.14**	**0.09**	**0.89**	**0.68**	**0.65**	**0.93**	**0.70**	0.50	**0.41**
ko→en	Online	0.51	0.37	5.16	4.29	3.86	4.71	3.89	1.61	2.28
	Gemma 3 4B	0.61	0.48	4.11	3.09	2.71	3.08	2.39	1.41	1.74
	DeepSeek-V3.2	0.38	0.28	2.88	1.99	1.61	2.21	1.49	0.82	1.13
	Claude 4.6	0.34	0.23	2.59	1.78	1.41	1.96	1.34	0.72	1.01
	Claude 4.8	0.26	0.17	1.94	1.31	1.14	1.56	1.04	0.56	0.77
	GPT-5.5	0.22	0.15	1.66	1.12	0.98	1.39	1.11	0.60	0.71
	DRT	0.20	0.13	1.32	0.94	0.88	1.24	0.92	**0.54**	0.57
	TransAgent	**0.18**	**0.12**	**1.22**	**0.86**	**0.81**	**1.14**	**0.84**	**0.48**	**0.52**
	SMART	**0.17**	**0.10**	**1.01**	**0.78**	**0.74**	**1.06**	**0.81**	0.59	**0.47**
it→en	Online	0.40	0.28	4.01	3.29	2.91	3.11	2.49	1.29	1.66
	Gemma 3 4B	0.49	0.35	3.41	2.52	2.18	2.44	1.86	1.11	1.39
	DeepSeek-V3.2	0.31	0.20	2.34	1.59	1.32	1.76	1.18	0.67	0.91
	Claude 4.6	0.27	0.17	2.14	1.46	1.18	1.56	1.08	0.58	0.82
	Claude 4.8	0.20	0.12	1.59	1.04	0.94	1.26	0.84	0.46	0.62
	GPT-5.5	0.18	0.11	1.36	0.90	0.82	1.14	0.87	0.48	0.57
	DRT	0.16	0.09	1.10	0.77	0.72	1.02	0.74	**0.43**	0.46
	TransAgent	**0.15**	**0.08**	**1.00**	**0.68**	**0.66**	**0.93**	**0.67**	**0.39**	**0.42**
	SMART	**0.14**	**0.07**	**0.84**	**0.62**	**0.60**	**0.86**	**0.64**	0.46	**0.38**
es-ES→en	Online	0.41	0.29	3.91	3.19	2.81	3.01	2.39	1.24	1.61
	Gemma 3 4B	0.51	0.36	3.31	2.42	2.11	2.36	1.79	1.07	1.35
	DeepSeek-V3.2	0.32	0.21	2.28	1.52	1.26	1.70	1.12	0.64	0.87
	Claude 4.6	0.28	0.18	2.09	1.41	1.14	1.52	1.04	0.56	0.80
	Claude 4.8	0.21	0.13	1.54	1.00	0.90	1.22	0.81	0.44	0.60
	GPT-5.5	0.19	0.11	1.32	0.86	0.78	1.10	0.84	0.46	0.55
	DRT	0.17	0.10	1.06	0.74	0.69	0.98	0.71	**0.41**	0.45
	TransAgent	**0.15**	**0.09**	**0.96**	**0.65**	**0.63**	**0.89**	**0.64**	**0.37**	**0.40**
	SMART	**0.14**	**0.07**	**0.80**	**0.58**	**0.57**	**0.82**	**0.61**	0.44	**0.36**
fr→en	Online	0.42	0.30	4.11	3.39	3.01	3.21	2.59	1.34	1.72
	Gemma 3 4B	0.52	0.38	3.51	2.59	2.24	2.51	1.92	1.14	1.40
	DeepSeek-V3.2	0.33	0.22	2.41	1.64	1.36	1.81	1.22	0.68	0.91
	Claude 4.6	0.29	0.19	2.19	1.49	1.21	1.62	1.11	0.59	0.84
	Claude 4.8	0.22	0.14	1.64	1.08	0.96	1.31	0.87	0.47	0.64
	GPT-5.5	0.20	0.12	1.39	0.92	0.84	1.19	0.90	0.49	0.59
	DRT	0.17	0.10	1.12	0.79	0.74	1.05	0.77	**0.44**	0.48
	TransAgent	**0.16**	**0.09**	**1.02**	**0.70**	**0.68**	**0.96**	**0.69**	**0.40**	**0.43**
	SMART	**0.15**	**0.08**	**0.86**	**0.64**	**0.62**	**0.89**	**0.67**	0.48	**0.39**
pt-PT→en	Online	0.67	0.29	3.06	1.73	2.13	2.96	2.25	1.56	1.31
	Gemma 3 4B	0.57	0.23	2.68	1.39	1.72	2.58	1.83	1.26	1.09
	DeepSeek-V3.2	0.38	0.18	2.00	1.09	1.21	1.95	1.43	0.95	0.82
	Claude 4.6	0.31	0.14	1.72	1.01	0.94	1.52	1.11	0.81	0.68
	Claude 4.8	0.28	0.12	1.50	0.92	0.87	1.33	1.04	0.75	0.61
	GPT-5.5	0.26	0.11	1.33	0.84	0.80	1.19	0.95	0.70	0.55
	DRT	0.23	0.10	1.19	0.78	0.76	1.12	0.83	0.61	0.50
	TransAgent	**0.20**	**0.09**	**1.06**	**0.71**	**0.69**	**1.03**	**0.79**	**0.55**	**0.46**
	SMART	**0.18**	**0.08**	**0.97**	**0.62**	**0.61**	**0.92**	**0.69**	**0.49**	**0.41**
pt-BR→en	Online	0.40	0.18	2.31	1.13	1.42	2.06	1.43	0.95	0.90
	Gemma 3 4B	0.33	0.15	1.83	0.98	1.19	1.73	1.26	0.82	0.75
	DeepSeek-V3.2	0.25	0.12	1.33	0.72	0.81	1.33	0.90	0.68	0.55

Continued on next page

Table 13 (continued)

Dir.	Method	Terminology		Accuracy			Fluency			Overall
		NameInc	TermInc	MisTrans	UndTrans	OvrTrans	Coher	Natur	Vivid	
	Claude 4.6	0.20	0.11	1.25	0.64	0.76	1.11	0.77	0.54	0.49
	Claude 4.8	0.19	0.10	1.11	0.61	0.67	1.04	0.71	0.48	0.45
	GPT-5.5	0.18	0.09	0.97	0.57	0.63	0.97	0.67	0.43	0.41
	DRT	0.16	0.08	0.86	0.52	0.59	0.92	0.59	0.41	0.37
	TransAgent	**0.14**	**0.07**	**0.77**	**0.48**	**0.54**	**0.84**	**0.54**	**0.38**	**0.34**
	SMART	**0.13**	**0.06**	**0.71**	**0.44**	**0.49**	**0.74**	**0.51**	**0.35**	**0.31**
	Online	0.35	0.23	2.12	1.84	2.07	2.51	1.56	1.40	1.08
	Gemma 3 4B	0.30	0.20	1.71	1.44	1.61	1.99	1.25	1.13	0.86
	DeepSeek-V3.2	0.24	0.14	1.20	1.00	1.11	1.46	1.04	0.78	0.62
	Claude 4.6	0.19	0.12	1.11	0.79	0.95	1.14	0.84	0.65	0.53
no→en	Claude 4.8	0.18	0.11	1.01	0.75	0.84	1.04	0.76	0.57	0.48
	GPT-5.5	0.16	0.10	0.91	0.67	0.74	0.95	0.67	0.52	0.42
	DRT	0.15	0.09	0.83	0.62	0.67	0.89	0.63	0.46	0.39
	TransAgent	**0.14**	**0.08**	**0.75**	**0.55**	**0.61**	**0.85**	**0.60**	**0.41**	**0.36**
	SMART	**0.13**	**0.07**	**0.68**	**0.50**	**0.53**	**0.75**	**0.54**	**0.39**	**0.32**
	Online	0.38	0.15	2.03	1.59	1.27	2.08	1.44	1.35	0.92
	Gemma 3 4B	0.29	0.12	1.68	1.32	1.10	1.70	1.15	1.07	0.76
	DeepSeek-V3.2	0.21	0.11	1.16	1.00	0.91	1.34	0.90	0.74	0.57
	Claude 4.6	0.19	0.10	1.00	0.80	0.77	1.05	0.81	0.58	0.48
da→en	Claude 4.8	0.17	0.09	0.92	0.72	0.69	0.92	0.73	0.51	0.43
	GPT-5.5	0.15	0.08	0.85	0.67	0.64	0.80	0.64	0.47	0.39
	DRT	0.14	0.07	0.77	0.63	0.56	0.70	0.59	0.44	0.35
	TransAgent	**0.13**	**0.06**	**0.71**	**0.56**	**0.51**	**0.66**	**0.53**	**0.39**	**0.32**
	SMART	**0.11**	**0.05**	**0.64**	**0.50**	**0.48**	**0.60**	**0.46**	**0.36**	**0.29**
	Online	0.46	0.21	2.14	1.71	1.34	2.00	1.71	1.37	0.97
	Gemma 3 4B	0.38	0.17	1.66	1.50	1.15	1.80	1.36	1.07	0.81
	DeepSeek-V3.2	0.26	0.14	1.33	1.06	0.85	1.32	0.93	0.77	0.60
	Claude 4.6	0.21	0.11	1.15	0.86	0.72	1.16	0.77	0.65	0.51
nl→en	Claude 4.8	0.19	0.10	1.06	0.76	0.69	1.05	0.71	0.59	0.47
	GPT-5.5	0.17	0.09	0.99	0.70	0.60	0.93	0.65	0.53	0.42
	DRT	0.15	0.08	0.87	0.66	0.54	0.88	0.58	0.46	0.38
	TransAgent	**0.13**	**0.07**	**0.78**	**0.58**	**0.49**	**0.81**	**0.55**	**0.41**	**0.34**
	SMART	**0.12**	**0.06**	**0.70**	**0.52**	**0.47**	**0.71**	**0.52**	**0.36**	**0.31**
	Online	0.44	0.19	2.67	1.63	1.57	2.58	1.54	1.29	1.07
	Gemma 3 4B	0.36	0.17	2.11	1.41	1.22	2.27	1.22	1.04	0.87
	DeepSeek-V3.2	0.28	0.13	1.65	1.14	1.01	1.55	0.96	0.81	0.68
	Claude 4.6	0.24	0.12	1.39	0.89	0.84	1.27	0.86	0.68	0.57
ro→en	Claude 4.8	0.21	0.11	1.22	0.80	0.74	1.17	0.79	0.62	0.51
	GPT-5.5	0.19	0.10	1.08	0.76	0.67	1.09	0.74	0.56	0.46
	DRT	0.18	0.09	0.97	0.68	0.62	0.96	0.70	0.49	0.42
	TransAgent	**0.16**	**0.08**	**0.85**	**0.62**	**0.57**	**0.84**	**0.61**	**0.43**	**0.37**
	SMART	**0.14**	**0.07**	**0.80**	**0.58**	**0.54**	**0.77**	**0.54**	**0.39**	**0.34**
	Online	0.51	0.20	2.79	1.52	2.06	2.65	2.38	1.54	1.20
	Gemma 3 4B	0.41	0.18	2.18	1.29	1.62	2.24	1.83	1.36	0.98
	DeepSeek-V3.2	0.33	0.15	1.80	0.96	1.31	1.59	1.38	1.02	0.76
	Claude 4.6	0.29	0.13	1.48	0.90	1.04	1.35	1.19	0.85	0.64
sv→en	Claude 4.8	0.25	0.12	1.33	0.79	0.98	1.22	1.07	0.78	0.58
	GPT-5.5	0.22	0.11	1.19	0.75	0.88	1.13	0.93	0.69	0.53
	DRT	0.20	0.10	1.13	0.67	0.78	1.04	0.84	0.62	0.48
	TransAgent	**0.19**	**0.09**	**1.05**	**0.63**	**0.69**	**0.96**	**0.75**	**0.56**	**0.44**
	SMART	**0.17**	**0.08**	**0.92**	**0.59**	**0.61**	**0.88**	**0.66**	**0.52**	**0.39**
	Online	0.44	0.24	2.58	2.09	2.01	2.26	1.96	1.72	1.19
	Gemma 3 4B	0.37	0.20	2.17	1.76	1.70	2.03	1.60	1.52	1.02
	DeepSeek-V3.2	0.29	0.14	1.52	1.22	1.20	1.63	1.21	1.10	0.74
	Claude 4.6	0.24	0.12	1.42	1.08	0.93	1.35	1.04	0.87	0.63
tr→en	Claude 4.8	0.22	0.11	1.27	1.02	0.82	1.25	0.95	0.79	0.57
	GPT-5.5	0.20	0.10	1.15	0.93	0.74	1.16	0.83	0.72	0.52
	DRT	0.18	0.09	1.01	0.84	0.69	1.10	0.77	0.63	0.47
	TransAgent	**0.16**	**0.08**	**0.93**	**0.74**	**0.63**	**0.98**	**0.71**	**0.57**	**0.43**
	SMART	**0.14**	**0.07**	**0.83**	**0.66**	**0.58**	**0.86**	**0.64**	**0.51**	**0.38**
	Online	0.39	0.21	3.17	1.86	1.55	2.02	1.73	1.44	1.13
	Gemma 3 4B	0.35	0.18	2.44	1.45	1.22	1.61	1.46	1.22	0.90
	DeepSeek-V3.2	0.27	0.14	1.85	1.01	0.96	1.32	1.01	0.85	0.68
	Claude 4.6	0.23	0.13	1.53	0.94	0.83	1.17	0.92	0.73	0.59
es-MX→en	Claude 4.8	0.21	0.11	1.33	0.88	0.76	1.07	0.80	0.64	0.53
	GPT-5.5	0.19	0.10	1.16	0.83	0.69	0.95	0.76	0.57	0.48
	DRT	0.17	0.09	1.07	0.74	0.63	0.89	0.71	0.50	0.43
	TransAgent	**0.15**	**0.08**	**0.94**	**0.67**	**0.60**	**0.84**	**0.65**	**0.47**	**0.40**
	SMART	**0.13**	**0.07**	**0.88**	**0.60**	**0.55**	**0.79**	**0.61**	**0.43**	**0.36**

Abbreviations. Terminology (NameInc = Name Inconsistency; TermInc = Term Inconsistency). **Accuracy** (MisTrans = Mistranslation; UndTrans = Undertranslation; OvrTrans = Overtranslation). **Fluency** (Coher = Coherence; Natur = Naturalness; Vivid = Vividness).

Table 14: Fine-grained SubMQM results for translation from 15 source locales into English: Linguistic Conventions, Technical, Locale Conventions, and Audience Appropriateness. All values are penalties (lower is better). Best and second-best values within each language direction and metric are highlighted with blue-gray and warm-beige backgrounds, respectively. These results provide the full breakdown corresponding to Table 3.

Dir.	Method	Linguistic Conventions				Technical			Locale Conventions		Audience Appropriateness		Overall
		MisPunc	MisCap	Gram	Space	LineBrk	CharLim	LineLim	LocErr	LangDet	Profan	Formal	
zh→en	Online	2.16	1.19	1.71	0.81	0.49	0.31	**0.06**	0.18	0.03	0.28	0.15	2.17
	Gemma 3 4B	1.21	0.56	0.88	0.59	1.51	1.04	0.44	0.26	0.05	0.46	0.31	1.64
	DeepSeek-V3.2	0.64	0.32	0.46	0.24	1.41	0.99	0.24	0.15	**0.01**	0.25	0.10	1.07
	Claude 4.6	0.59	0.18	0.39	0.18	1.31	0.89	0.29	0.06	0.04	0.15	0.15	0.96
	Claude 4.8	0.42	0.11	0.28	0.12	1.11	0.72	0.22	0.03	0.03	**0.11**	0.12	0.73
	GPT-5.5	0.34	0.08	0.24	0.10	1.38	0.92	0.32	0.06	**0.00**	0.18	0.07	0.67
	DRT	0.28	0.05	0.20	0.07	0.51	0.26	0.13	**0.02**	0.03	0.16	**0.05**	0.55
	TransAgent	**0.24**	**0.04**	**0.18**	**0.06**	**0.44**	**0.20**	0.11	0.05	**0.00**	**0.10**	0.09	**0.50**
	SMART	**0.19**	**0.02**	**0.15**	**0.04**	**0.16**	**0.02**	**0.05**	**0.01**	0.03	0.14	**0.03**	**0.44**
de→en	Online	1.51	0.79	1.16	0.59	0.39	0.21	**0.09**	0.49	0.08	0.26	0.11	1.80
	Gemma 3 4B	1.01	0.46	0.74	0.46	1.41	0.92	0.38	0.32	**0.05**	0.39	0.24	1.49
	DeepSeek-V3.2	0.62	0.16	0.38	0.18	1.28	0.86	0.31	0.20	0.22	**0.08**	0.96	1.02
	Claude 4.6	0.52	0.14	0.32	0.14	1.21	0.79	0.28	0.16	**0.00**	0.14	0.12	0.88
	Claude 4.8	0.36	0.08	0.23	0.09	1.01	0.62	0.21	0.12	**0.00**	**0.10**	0.10	0.67
	GPT-5.5	0.30	0.06	0.20	0.07	1.24	0.82	0.28	0.10	**0.00**	0.16	0.06	0.62
	DRT	0.25	0.03	0.17	0.05	0.46	0.22	0.12	**0.08**	**0.00**	0.14	**0.04**	0.50
	TransAgent	**0.22**	**0.02**	**0.15**	**0.04**	**0.38**	**0.16**	0.10	**0.07**	**0.00**	**0.08**	0.08	**0.45**
	SMART	**0.17**	**0.00**	**0.13**	**0.03**	**0.14**	**0.01**	**0.07**	**0.07**	**0.00**	0.12	**0.02**	**0.41**
ko→en	Online	2.31	1.29	1.81	0.84	0.54	0.32	0.15	0.22	**0.02**	0.32	0.16	2.28
	Gemma 3 4B	1.26	0.62	0.94	0.64	1.56	1.09	0.46	0.29	0.05	0.50	0.34	1.74
	DeepSeek-V3.2	0.78	0.24	0.49	0.26	1.44	1.02	0.36	0.16	**0.00**	0.27	0.11	1.13
	Claude 4.6	0.62	0.20	0.42	0.20	1.36	0.92	0.32	0.13	**0.00**	0.17	0.15	1.01
	Claude 4.8	0.44	0.12	0.30	0.13	1.14	0.74	0.25	0.10	**0.00**	**0.13**	0.12	0.77
	GPT-5.5	0.36	0.09	0.26	0.11	1.42	0.96	0.34	0.09	**0.00**	0.20	0.07	0.71
	DRT	0.29	0.06	0.22	0.08	0.52	0.28	0.14	**0.08**	**0.00**	0.18	**0.05**	0.57
	TransAgent	**0.25**	**0.04**	**0.19**	**0.07**	**0.45**	**0.22**	**0.12**	**0.07**	**0.00**	**0.11**	0.10	**0.52**
	SMART	**0.21**	**0.02**	**0.16**	**0.05**	**0.19**	**0.01**	**0.09**	**0.07**	**0.00**	0.16	**0.03**	**0.47**
it→en	Online	1.41	0.69	1.04	0.52	**0.36**	**0.14**	**0.10**	0.39	0.04	0.23	0.09	1.66
	Gemma 3 4B	0.94	0.42	0.68	0.42	1.34	0.86	0.34	0.26	**0.03**	0.36	0.20	1.39
	DeepSeek-V3.2	0.57	0.14	0.37	0.14	1.22	0.80	0.28	0.18	**0.00**	0.21	0.07	0.91
	Claude 4.6	0.48	0.12	0.29	0.12	1.16	0.74	0.26	0.14	**0.00**	0.13	0.11	0.82
	Claude 4.8	0.33	0.06	0.21	0.08	0.96	0.58	0.20	0.10	**0.00**	**0.09**	0.09	0.62
	GPT-5.5	0.27	0.04	0.18	0.06	1.18	0.76	0.26	0.09	**0.00**	0.16	0.04	0.57
	DRT	0.23	0.02	0.15	0.04	0.44	0.20	0.11	**0.07**	**0.00**	0.14	**0.02**	0.46
	TransAgent	**0.20**	**0.01**	**0.14**	**0.03**	**0.36**	**0.14**	**0.10**	**0.06**	**0.00**	**0.07**	0.07	**0.42**
	SMART	**0.16**	**0.00**	**0.12**	**0.02**	**0.15**	**0.00**	**0.07**	**0.06**	**0.00**	0.12	**0.00**	**0.38**
es-ES→en	Online	1.36	0.64	0.98	0.49	**0.34**	**0.12**	0.10	0.36	0.03	0.22	0.08	1.61
	Gemma 3 4B	0.91	0.39	0.64	0.39	1.31	0.84	0.32	0.24	**0.02**	0.34	0.18	1.35
	DeepSeek-V3.2	0.54	0.12	0.34	0.12	1.20	0.78	0.27	0.16	**0.00**	0.20	0.06	0.87
	Claude 4.6	0.46	0.11	0.28	0.11	1.14	0.72	0.25	0.13	**0.00**	0.12	0.10	0.80
	Claude 4.8	0.32	0.06	0.20	0.07	0.94	0.56	0.19	0.09	**0.00**	**0.09**	0.08	0.60
	GPT-5.5	0.26	0.04	0.17	0.05	1.16	0.74	0.25	0.08	**0.00**	0.15	0.03	0.55
	DRT	0.22	0.02	0.14	0.03	0.42	0.18	0.11	**0.07**	**0.00**	0.13	**0.01**	0.45
	TransAgent	**0.19**	**0.01**	**0.13**	**0.02**	**0.34**	**0.12**	**0.09**	**0.06**	**0.00**	**0.06**	0.06	**0.40**
	SMART	**0.15**	**0.00**	**0.11**	**0.01**	**0.14**	**0.00**	**0.06**	**0.06**	**0.00**	0.11	**0.00**	**0.36**
fr→en	Online	1.46	0.74	1.08	0.54	0.38	0.16	**0.11**	0.42	0.05	0.24	0.10	1.72
	Gemma 3 4B	0.96	0.44	0.71	0.44	1.36	0.89	0.36	0.28	**0.04**	0.38	0.22	1.40
	DeepSeek-V3.2	0.58	0.14	0.38	0.14	1.24	0.82	0.30	0.19	**0.00**	0.22	0.07	0.91
	Claude 4.6	0.49	0.11	0.30	0.11	1.18	0.76	0.27	0.15	**0.00**	0.13	0.11	0.84
	Claude 4.8	0.34	0.05	0.22	0.07	0.98	0.60	0.21	0.11	**0.00**	**0.10**	0.09	0.64
	GPT-5.5	0.28	0.03	0.19	0.05	1.21	0.79	0.27	0.10	**0.00**	0.16	0.04	0.59
	DRT	0.24	**0.01**	0.16	0.03	0.45	0.21	0.13	**0.08**	**0.00**	0.14	**0.02**	0.48
	TransAgent	**0.21**	**0.00**	**0.14**	**0.02**	**0.37**	**0.15**	**0.11**	**0.07**	**0.00**	**0.07**	0.07	**0.43**
	SMART	**0.17**	**0.00**	**0.12**	**0.01**	**0.15**	**0.01**	**0.08**	**0.07**	**0.00**	0.13	**0.00**	**0.39**
pt-PT→en	Online	0.60	0.08	0.39	0.09	0.47	0.08	0.20	0.22	0.09	0.36	0.09	1.31
	Gemma 3 4B	0.52	0.07	0.31	0.08	0.38	0.07	0.16	0.19	0.08	0.31	0.08	1.09
	DeepSeek-V3.2	0.37	0.06	0.25	0.07	0.27	0.06	0.13	0.13	0.07	0.24	0.07	0.82
	Claude 4.6	0.29	0.05	0.22	0.06	0.25	0.05	0.12	0.12	0.06	0.21	0.06	0.68
	Claude 4.8	0.27	0.04	0.19	0.05	0.23	0.04	0.11	0.11	0.05	0.19	0.05	0.61
	GPT-5.5	0.24	0.03	0.18	0.04	0.21	0.03	0.10	0.10	0.04	0.18	0.04	0.55
	DRT	0.21	0.02	0.16	0.03	0.19	0.02	0.09	0.09	0.03	0.17	0.03	0.50
	TransAgent	**0.19**	**0.01**	**0.15**	**0.02**	**0.18**	**0.01**	**0.08**	**0.08**	**0.02**	**0.16**	**0.02**	**0.46**
	SMART	**0.17**	**0.00**	**0.14**	**0.01**	**0.17**	**0.00**	**0.07**	**0.07**	**0.01**	**0.14**	**0.01**	**0.41**
pt-BR→en	Online	0.38	0.08	0.23	0.09	0.40	0.08	0.18	0.14	0.09	0.31	0.09	0.90
	Gemma 3 4B	0.34	0.07	0.19	0.08	0.31	0.07	0.15	0.12	0.08	0.26	0.08	0.75
	DeepSeek-V3.2	0.25	0.06	0.15	0.07	0.23	0.06	0.11	0.11	0.07	0.20	0.07	0.55
	Claude 4.6	0.22	0.05	0.14	0.06	0.19	0.05	0.10	0.10	0.06	0.17	0.06	0.49
	Claude 4.8	0.20	0.04	0.13	0.05	0.18	0.04	0.09	0.09	0.05	0.16	0.05	0.45
	GPT-5.5	0.18	0.03	0.12	0.04	0.17	0.03	0.08	0.08	0.04	0.14	0.04	0.41
	DRT	0.16	0.02	0.11	0.03	0.15	0.02	0.07	0.07	0.03	0.12	0.03	0.37
	TransAgent	**0.14**	**0.01**	**0.10**	**0.02**	**0.13**	**0.01**	**0.06**	**0.06**	**0.02**	**0.11**	**0.02**	**0.34**
	SMART	**0.13**	**0.00**	**0.09**	**0.01**	**0.12**	**0.00**	**0.05**	**0.05**	**0.01**	**0.10**	**0.01**	**0.31**
no→en	Online	0.33	0.09	0.36	0.09	0.40	0.09	0.13	0.14	0.09	0.28	0.09	1.08

Continued on next page

Table 14 (continued)

Dir.	Method	Linguistic Conventions				Technical			Locale Conventions		Audience Appropriateness		Overall
		MisPunc	MisCap	Gram	Space	LineBrk	CharLim	LineLim	LocErr	LangDet	Profan	Formal	
	Gemma 3 4B	0.28	0.08	0.28	0.08	0.32	0.08	0.12	0.12	0.08	0.25	0.08	0.86
	DeepSeek-V3.2	0.23	0.07	0.19	0.07	0.23	0.07	0.11	0.11	0.07	0.20	0.07	0.62
	Claude 4.6	0.20	0.06	0.16	0.06	0.21	0.06	0.10	0.10	0.06	0.17	0.06	0.53
	Claude 4.8	0.19	0.05	0.14	0.05	0.19	0.05	0.09	0.09	0.05	0.15	0.05	0.48
	GPT-5.5	0.17	0.04	0.12	0.04	0.18	0.04	0.08	0.08	0.04	0.13	0.04	0.42
	DRT	0.15	0.03	0.11	0.03	0.16	0.03	0.07	0.07	0.03	0.12	0.03	0.39
	TransAgent	**0.14**	**0.02**	**0.10**	**0.02**	**0.15**	**0.02**	**0.06**	**0.06**	**0.02**	**0.11**	**0.02**	**0.36**
	SMART	**0.13**	**0.01**	**0.09**	**0.01**	**0.13**	**0.01**	**0.05**	**0.05**	**0.00**	**0.10**	**0.01**	**0.32**
	Online	0.38	0.09	0.24	0.09	0.30	0.09	0.21	0.18	0.09	0.26	0.08	0.92
	Gemma 3 4B	0.30	0.08	0.20	0.08	0.25	0.08	0.16	0.16	0.08	0.22	0.07	0.76
	DeepSeek-V3.2	0.22	0.07	0.16	0.07	0.18	0.07	0.11	0.11	0.07	0.15	0.06	0.57
	Claude 4.6	0.19	0.06	0.15	0.06	0.17	0.06	0.10	0.10	0.06	0.14	0.05	0.48
da→en	Claude 4.8	0.17	0.05	0.14	0.05	0.16	0.05	0.09	0.09	0.05	0.13	0.04	0.43
	GPT-5.5	0.15	0.04	0.12	0.04	0.14	0.04	0.08	0.08	0.04	0.12	0.03	0.39
	DRT	0.14	0.03	0.11	0.03	0.13	0.03	0.07	0.07	0.03	0.11	0.02	0.35
	TransAgent	**0.13**	**0.02**	**0.10**	**0.02**	**0.12**	**0.02**	**0.06**	**0.06**	**0.02**	**0.10**	**0.01**	**0.32**
	SMART	**0.12**	**0.01**	**0.09**	**0.01**	**0.11**	**0.01**	**0.05**	**0.05**	**0.00**	**0.09**	**0.00**	**0.29**
	Online	0.42	0.09	0.25	0.09	0.34	0.09	0.14	0.17	0.08	0.23	0.09	0.97
	Gemma 3 4B	0.34	0.08	0.22	0.08	0.28	0.08	0.12	0.14	0.07	0.19	0.08	0.81
	DeepSeek-V3.2	0.23	0.07	0.18	0.07	0.21	0.07	0.11	0.11	0.06	0.15	0.07	0.60
	Claude 4.6	0.20	0.06	0.15	0.06	0.19	0.06	0.10	0.10	0.05	0.14	0.06	0.51
nl→en	Claude 4.8	0.19	0.05	0.14	0.05	0.18	0.05	0.09	0.09	0.04	0.13	0.05	0.47
	GPT-5.5	0.18	0.04	0.12	0.04	0.16	0.04	0.08	0.08	0.03	0.12	0.04	0.42
	DRT	0.16	0.03	0.11	0.03	0.15	0.03	0.07	0.07	0.02	0.11	0.03	0.38
	TransAgent	**0.14**	**0.02**	**0.10**	**0.02**	**0.13**	**0.02**	**0.06**	**0.06**	**0.01**	**0.10**	**0.02**	**0.34**
	SMART	**0.12**	**0.01**	**0.09**	**0.01**	**0.12**	**0.01**	**0.05**	**0.05**	**0.00**	**0.09**	**0.01**	**0.31**
	Online	0.42	0.08	0.25	0.09	0.40	0.08	0.23	0.17	0.09	0.31	0.08	1.07
	Gemma 3 4B	0.33	0.07	0.22	0.08	0.31	0.07	0.19	0.14	0.08	0.25	0.07	0.87
	DeepSeek-V3.2	0.23	0.06	0.15	0.07	0.24	0.06	0.13	0.11	0.07	0.17	0.06	0.68
	Claude 4.6	0.20	0.05	0.14	0.06	0.20	0.05	0.10	0.10	0.06	0.15	0.05	0.57
ro→en	Claude 4.8	0.19	0.04	0.13	0.05	0.18	0.04	0.09	0.09	0.05	0.14	0.04	0.51
	GPT-5.5	0.18	0.03	0.12	0.04	0.17	0.03	0.08	0.08	0.04	0.13	0.03	0.46
	DRT	0.16	0.02	0.11	0.03	0.16	0.02	0.07	0.07	0.03	0.12	0.02	0.42
	TransAgent	**0.14**	**0.01**	**0.10**	**0.02**	**0.15**	**0.01**	**0.06**	**0.06**	**0.02**	**0.11**	**0.01**	**0.37**
	SMART	**0.13**	**0.00**	**0.09**	**0.01**	**0.14**	**0.00**	**0.05**	**0.05**	**0.01**	**0.10**	**0.00**	**0.34**
	Online	0.50	0.09	0.30	0.09	0.41	0.08	0.20	0.20	0.09	0.31	0.09	1.20
	Gemma 3 4B	0.45	0.08	0.27	0.08	0.32	0.07	0.16	0.17	0.08	0.25	0.08	0.98
	DeepSeek-V3.2	0.31	0.07	0.21	0.07	0.25	0.06	0.12	0.13	0.07	0.19	0.07	0.76
	Claude 4.6	0.24	0.06	0.19	0.06	0.24	0.05	0.11	0.12	0.06	0.18	0.06	0.64
sv→en	Claude 4.8	0.22	0.05	0.17	0.05	0.22	0.04	0.10	0.11	0.05	0.16	0.05	0.58
	GPT-5.5	0.20	0.04	0.16	0.04	0.20	0.03	0.09	0.10	0.04	0.15	0.04	0.53
	DRT	0.18	0.03	0.15	0.03	0.18	0.02	0.08	0.09	0.03	0.13	0.03	0.48
	TransAgent	**0.17**	**0.02**	**0.14**	**0.02**	**0.17**	**0.01**	**0.07**	**0.08**	**0.02**	**0.12**	**0.02**	**0.44**
	SMART	**0.16**	**0.01**	**0.13**	**0.01**	**0.16**	**0.00**	**0.06**	**0.07**	**0.01**	**0.11**	**0.01**	**0.39**
	Online	0.42	0.09	0.37	0.09	0.39	0.09	0.16	0.27	0.09	0.32	0.09	1.19
	Gemma 3 4B	0.37	0.08	0.31	0.08	0.31	0.08	0.14	0.21	0.08	0.28	0.08	1.02
	DeepSeek-V3.2	0.27	0.07	0.23	0.07	0.23	0.07	0.12	0.15	0.07	0.19	0.07	0.74
	Claude 4.6	0.22	0.06	0.19	0.06	0.21	0.06	0.11	0.13	0.06	0.18	0.06	0.63
tr→en	Claude 4.8	0.21	0.05	0.18	0.05	0.19	0.05	0.10	0.11	0.05	0.17	0.05	0.57
	GPT-5.5	0.20	0.04	0.17	0.04	0.17	0.04	0.09	0.10	0.04	0.15	0.04	0.52
	DRT	0.19	0.03	0.16	0.03	0.16	0.03	0.08	0.09	0.03	0.13	0.03	0.47
	TransAgent	**0.17**	**0.02**	**0.15**	**0.02**	**0.15**	**0.02**	**0.07**	**0.08**	**0.02**	**0.12**	**0.02**	**0.43**
	SMART	**0.15**	**0.01**	**0.13**	**0.01**	**0.14**	**0.01**	**0.06**	**0.07**	**0.01**	**0.11**	**0.01**	**0.38**
	Online	0.39	0.08	0.27	0.09	0.41	0.09	0.18	0.26	0.09	0.41	0.09	1.13
	Gemma 3 4B	0.33	0.07	0.24	0.08	0.35	0.08	0.14	0.21	0.08	0.34	0.08	0.90
	DeepSeek-V3.2	0.26	0.06	0.19	0.07	0.26	0.07	0.12	0.16	0.07	0.23	0.07	0.68
	Claude 4.6	0.24	0.05	0.18	0.06	0.24	0.06	0.11	0.12	0.06	0.19	0.06	0.59
es-MX→en	Claude 4.8	0.22	0.04	0.17	0.05	0.21	0.05	0.10	0.11	0.05	0.17	0.05	0.53
	GPT-5.5	0.21	0.03	0.15	0.04	0.19	0.04	0.09	0.10	0.04	0.16	0.04	0.48
	DRT	0.19	0.02	0.14	0.03	0.17	0.03	0.08	0.09	0.03	0.15	0.03	0.43
	TransAgent	**0.17**	**0.01**	**0.13**	**0.02**	**0.15**	**0.02**	**0.07**	**0.08**	**0.02**	**0.14**	**0.02**	**0.40**
	SMART	**0.15**	**0.00**	**0.12**	**0.01**	**0.14**	**0.00**	**0.06**	**0.07**	**0.01**	**0.13**	**0.00**	**0.36**

Abbreviations. Linguistic Conventions (MisPunc = Mispunctuation; MisCap = Miscapitalization; Gram = Grammar; Space = Spacing Error). **Technical** (LineBrk = Incorrect Line Breaking; CharLim = Exceeding Characters per Line; LineLim = Exceeding Lines per Box). **Locale Conventions** (LocErr = Localization Error; LangDet = Language Detection Error). **Audience Appropriateness** (Profan = Profanity; Formal = Formality Error).

F.3 Backbone and Judge Analysis

Tables 15 and 16 report the complete English→locale backbone and judge analysis, while Tables 17 and 18 report the corresponding locale→English results. These tables are the full fine-grained counterpart of the representative results summarized in Table 5. Across directions, applying SMART to Gemma 3 4B or DeepSeek-V3.2 reduces the Overall penalty relative to direct translation with the same model, indicating that the gain is not explained solely by the strength of Claude Sonnet 4.6.

Changing only the evaluator from Claude Sonnet 4.6 to GPT-5.5 produces substantially smaller differences than changing the translation scaffold or backbone. For example, on en→zh the Overall

penalty changes from 0.48 to 0.47, while direct Gemma 3 4B and SMART with Gemma 3 4B differ by 1.46 versus 0.67. The same pattern appears in the reverse direction, where zh→en changes from 0.44 to 0.44 under the alternative judge, compared with 1.64 versus 0.62 for direct and SMART-wrapped Gemma 3 4B. This separates the effect of the translation framework from small evaluator-specific score variation.

Table 15: Fine-grained backbone and judge analysis on Subtitle Arena for translation from English into 15 target locales: Terminology, Accuracy, and Fluency. All values are penalties (lower is better). Best and second-best values within each language direction and metric are highlighted with blue-gray and warm-beige backgrounds, respectively. These results provide the full breakdown corresponding to Table 5.

Dir.	Method / Setting	Terminology		Accuracy			Fluency			Overall
		NameInc	TermInc	MisTrans	UndTrans	OvrTrans	Coher	Natur	Vivid	
en→zh	Gemma 3 4B	0.52	0.46	3.45	2.68	2.15	2.65	2.12	1.15	1.46
	DeepSeek-V3.2	0.31	0.25	2.52	1.76	1.30	1.88	1.29	**0.58**	0.93
	SMART (Gemma 3 4B)	0.30	0.16	1.20	1.08	1.14	1.39	1.28	1.14	0.67
	SMART (DeepSeek-V3.2)	**0.22**	**0.13**	0.96	0.88	0.84	1.14	1.00	0.98	0.53
	SMART (Judge: GPT-5.5)	**0.18**	**0.12**	**0.90**	**0.83**	**0.68**	**1.03**	**0.89**	**0.78**	**0.47**
	SMART (Claude 4.6)	**0.18**	**0.12**	**0.86**	**0.81**	**0.73**	**1.08**	**0.94**	0.83	**0.48**
en→de	Gemma 3 4B	0.78	0.62	4.85	4.12	3.56	3.45	2.89	1.88	2.21
	DeepSeek-V3.2	0.50	0.37	3.52	2.81	2.38	2.54	1.98	**1.25**	1.50
	SMART (Gemma 3 4B)	0.34	0.32	1.80	1.68	1.55	2.46	2.28	1.80	1.23
	SMART (DeepSeek-V3.2)	**0.30**	0.23	1.54	1.50	1.19	1.87	1.76	1.42	**0.99**
	SMART (Judge: GPT-5.5)	**0.29**	**0.20**	**1.43**	**1.23**	**1.00**	**1.65**	**1.48**	1.39	**0.87**
	SMART (Claude 4.6)	**0.29**	**0.21**	**1.39**	**1.28**	**1.05**	**1.62**	**1.46**	**1.30**	**0.87**
en→ko	Gemma 3 4B	0.75	0.61	3.82	3.12	2.74	2.98	2.54	1.78	1.82
	DeepSeek-V3.2	0.48	0.35	2.58	1.98	1.72	2.08	1.66	**1.14**	1.17
	SMART (Gemma 3 4B)	0.45	0.38	2.16	2.06	1.86	1.98	2.12	1.77	1.26
	SMART (DeepSeek-V3.2)	0.38	**0.28**	1.93	1.68	1.44	1.70	1.68	1.58	1.03
	SMART (Judge: GPT-5.5)	**0.36**	**0.24**	**1.79**	**1.51**	**1.28**	**1.60**	**1.36**	**1.32**	**0.91**
	SMART (Claude 4.6)	**0.34**	**0.24**	**1.76**	**1.56**	**1.33**	**1.56**	**1.41**	**1.32**	**0.92**
en→it	Gemma 3 4B	0.78	0.62	4.95	4.15	3.65	3.98	3.42	2.35	2.34
	DeepSeek-V3.2	0.50	0.36	3.60	2.79	2.44	3.08	2.59	**1.71**	1.62
	SMART (Gemma 3 4B)	0.40	0.25	2.16	1.86	1.89	3.25	2.99	2.84	1.56
	SMART (DeepSeek-V3.2)	0.30	0.19	1.89	1.55	1.62	2.76	2.50	2.23	**1.30**
	SMART (Judge: GPT-5.5)	**0.28**	**0.17**	**1.61**	**1.45**	**1.30**	**2.34**	**2.07**	2.16	**1.13**
	SMART (Claude 4.6)	**0.26**	**0.16**	**1.66**	**1.46**	**1.35**	**2.36**	**2.12**	**1.97**	**1.13**
en→es-ES	Gemma 3 4B	0.82	0.65	5.12	4.28	3.76	3.82	3.28	2.25	2.36
	DeepSeek-V3.2	0.52	0.38	3.70	2.89	2.51	2.86	2.38	**1.60**	1.62
	SMART (Gemma 3 4B)	0.69	0.52	2.76	2.78	2.25	3.13	2.18	2.27	1.71
	SMART (DeepSeek-V3.2)	0.50	0.38	2.38	2.05	1.69	2.51	1.93	1.88	1.36
	SMART (Judge: GPT-5.5)	**0.46**	**0.32**	**1.91**	**1.71**	**1.54**	**2.01**	**1.75**	1.77	**1.17**
	SMART (Claude 4.6)	**0.44**	**0.34**	**1.96**	**1.76**	**1.59**	**2.06**	**1.80**	**1.66**	**1.18**
en→fr	Gemma 3 4B	0.72	0.56	4.90	4.08	3.58	3.95	3.38	2.32	2.31
	DeepSeek-V3.2	0.42	0.28	3.55	2.80	2.40	3.04	2.50	**1.69**	1.60
	SMART (Gemma 3 4B)	0.33	0.13	2.86	1.99	1.94	3.84	3.17	2.90	1.74
	SMART (DeepSeek-V3.2)	**0.25**	**0.11**	2.22	1.72	1.69	2.87	2.60	2.33	**1.40**
	SMART (Judge: GPT-5.5)	**0.20**	**0.10**	**1.80**	**1.53**	**1.38**	**2.52**	**2.30**	**1.92**	**1.19**
	SMART (Claude 4.6)	**0.20**	**0.10**	**1.82**	**1.58**	**1.43**	**2.41**	**2.13**	1.97	**1.19**
en→pt-PT	Gemma 3 4B	1.38	1.04	6.80	5.06	5.47	6.44	4.75	5.14	3.72
	DeepSeek-V3.2	1.01	0.84	5.05	3.62	3.70	4.73	3.70	4.07	2.73
	SMART (Gemma 3 4B)	0.76	0.49	3.91	2.48	2.20	2.68	2.88	2.90	1.89
	SMART (DeepSeek-V3.2)	**0.60**	0.42	2.81	2.24	1.87	**2.43**	**2.17**	2.08	1.50
	SMART (Judge: GPT-5.5)	**0.53**	**0.35**	**2.21**	**1.76**	**1.67**	**2.38**	**1.99**	**2.00**	**1.30**
	SMART (Claude 4.6)	**0.53**	**0.37**	**2.41**	**1.93**	**1.83**	**2.43**	2.18	**2.03**	**1.39**
en→pt-BR	Gemma 3 4B	1.04	0.78	4.24	3.72	3.26	4.30	3.69	3.60	2.56
	DeepSeek-V3.2	0.71	0.53	3.27	2.88	2.52	3.29	2.73	2.52	1.91
	SMART (Gemma 3 4B)	0.53	0.38	2.79	1.80	1.52	2.68	2.17	2.24	1.42
	SMART (DeepSeek-V3.2)	0.39	0.34	2.03	1.61	1.36	2.17	**1.68**	1.68	1.14
	SMART (Judge: GPT-5.5)	**0.31**	**0.30**	**1.69**	**1.40**	**1.25**	**1.71**	**1.53**	**1.32**	**0.97**
	SMART (Claude 4.6)	**0.36**	**0.32**	**1.79**	**1.48**	**1.26**	**1.85**	1.69	**1.45**	**1.04**
en→no	Gemma 3 4B	1.30	0.97	4.32	4.99	3.96	5.22	5.42	3.57	2.99
	DeepSeek-V3.2	0.93	0.76	3.33	3.37	3.04	3.96	3.64	2.54	2.18
	SMART (Gemma 3 4B)	0.58	0.41	2.29	2.49	2.32	3.07	2.60	2.09	1.60
	SMART (DeepSeek-V3.2)	0.49	0.35	2.02	1.93	1.96	2.26	2.10	1.65	1.29
	SMART (Judge: GPT-5.5)	**0.43**	**0.31**	**1.64**	**1.70**	**1.59**	**1.90**	**1.82**	**1.39**	**1.10**
	SMART (Claude 4.6)	**0.46**	**0.33**	**1.78**	**1.86**	**1.74**	**2.05**	**1.99**	**1.53**	**1.19**
en→da	Gemma 3 4B	1.10	0.88	6.12	4.72	3.83	6.45	5.29	4.23	3.30
	DeepSeek-V3.2	0.84	0.60	4.31	3.22	3.00	4.70	3.56	2.93	2.34
	SMART (Gemma 3 4B)	0.63	0.50	3.23	2.55	2.62	2.84	2.61	2.51	1.79

Continued on next page

Table 15 (continued)

Dir.	Method / Setting	Terminology		Accuracy			Fluency			Overall
		NameInc	TermInc	MisTrans	UndTrans	OvrTrans	Coher	Natur	Vivid	
	SMART (DeepSeek-V3.2)	0.55	**0.39**	2.54	2.02	1.88	2.39	1.92	1.89	1.39
	SMART (Judge: GPT-5.5)	**0.52**	**0.34**	**2.11**	**1.72**	**1.65**	**2.09**	**1.64**	**1.47**	**1.19**
	SMART (Claude 4.6)	**0.50**	**0.34**	**2.30**	**1.88**	**1.74**	**2.12**	**1.80**	**1.61**	**1.26**
en→nl	Gemma 3 4B	1.21	1.08	5.75	4.02	5.25	5.54	5.17	5.14	3.41
	DeepSeek-V3.2	0.82	0.77	4.67	3.17	3.74	4.17	3.99	3.71	2.57
	SMART (Gemma 3 4B)	0.66	0.56	3.74	2.22	2.62	3.01	2.46	2.63	1.80
	SMART (DeepSeek-V3.2)	0.59	0.46	2.73	1.92	2.19	2.47	2.17	2.20	1.48
	SMART (Judge: GPT-5.5)	**0.46**	**0.41**	**2.17**	**1.85**	**1.72**	**2.30**	**2.08**	**1.89**	**1.29**
	SMART (Claude 4.6)	**0.52**	**0.40**	**2.36**	**1.91**	**1.88**	**2.44**	**2.14**	**2.02**	**1.38**
en→ro	Gemma 3 4B	1.12	0.94	5.01	4.20	4.23	5.61	3.79	4.67	2.96
	DeepSeek-V3.2	0.81	0.65	3.50	3.13	3.24	3.80	2.95	3.51	2.17
	SMART (Gemma 3 4B)	0.55	0.48	2.39	2.08	1.92	2.53	2.04	2.23	1.45
	SMART (DeepSeek-V3.2)	**0.46**	0.37	1.86	1.81	**1.56**	1.88	1.49	1.66	1.14
	SMART (Judge: GPT-5.5)	**0.43**	**0.32**	**1.68**	**1.44**	**1.42**	**1.80**	**1.34**	**1.48**	**1.01**
	SMART (Claude 4.6)	**0.43**	**0.33**	**1.84**	**1.58**	**1.56**	**1.84**	**1.46**	**1.51**	**1.07**
en→sv	Gemma 3 4B	0.93	0.78	3.86	4.33	3.62	3.99	3.16	4.01	2.57
	DeepSeek-V3.2	0.77	0.56	3.14	2.91	2.59	3.31	2.44	2.75	1.91
	SMART (Gemma 3 4B)	0.48	0.36	2.76	2.19	1.68	2.50	2.27	1.79	1.43
	SMART (DeepSeek-V3.2)	0.37	0.32	2.03	1.72	1.47	2.07	1.65	1.55	1.14
	SMART (Judge: GPT-5.5)	**0.32**	**0.25**	**1.66**	**1.46**	**1.21**	**1.61**	**1.48**	**1.23**	**0.95**
	SMART (Claude 4.6)	**0.36**	**0.28**	**1.81**	**1.60**	**1.34**	**1.76**	**1.61**	**1.36**	**1.04**
en→tr	Gemma 3 4B	1.18	0.85	5.78	4.39	3.64	5.10	4.48	4.30	3.04
	DeepSeek-V3.2	0.93	0.62	4.18	3.14	2.58	3.94	3.46	3.45	2.26
	SMART (Gemma 3 4B)	0.57	0.52	2.77	2.76	1.92	2.43	2.50	2.76	1.66
	SMART (DeepSeek-V3.2)	0.47	0.39	2.30	2.15	1.71	2.12	2.09	2.04	1.35
	SMART (Judge: GPT-5.5)	**0.41**	**0.37**	**2.01**	**1.82**	**1.41**	**1.85**	**1.84**	**1.78**	**1.17**
	SMART (Claude 4.6)	**0.45**	**0.38**	**2.19**	**1.98**	**1.55**	**2.01**	**1.83**	**1.80**	**1.25**
en→es-MX	Gemma 3 4B	1.17	1.00	5.30	6.87	4.85	5.40	4.58	4.67	3.43
	DeepSeek-V3.2	0.94	0.69	4.05	4.73	3.28	3.72	3.39	3.86	2.51
	SMART (Gemma 3 4B)	0.67	0.50	2.50	2.40	2.31	3.24	2.85	2.53	1.72
	SMART (DeepSeek-V3.2)	0.52	0.41	**2.15**	2.14	1.86	2.38	2.13	2.10	1.39
	SMART (Judge: GPT-5.5)	**0.45**	**0.36**	**2.09**	**1.88**	**1.58**	**1.96**	**1.97**	**1.81**	**1.23**
	SMART (Claude 4.6)	**0.50**	**0.38**	**2.15**	**2.05**	**1.73**	**2.08**	**2.01**	**1.97**	**1.31**

Abbreviations. Terminology (NameInc = Name Inconsistency; TermInc = Term Inconsistency). **Accuracy** (MisTrans = Mistranslation; UndTrans = Undertranslation; OvrTrans = Overtranslation). **Fluency** (Coher = Coherence; Natur = Naturalness; Vivid = Vividness).

Table 16: Fine-grained backbone and judge analysis on Subtitle Arena for translation from English into 15 target locales: Linguistic Conventions, Technical, Locale Conventions, and Audience Appropriateness. All values are penalties (lower is better). Best and second-best values within each language direction and metric are highlighted with blue-gray and warm-beige backgrounds, respectively. These results provide the full breakdown corresponding to Table 5.

Dir.	Method / Setting	Linguistic Conventions				Technical			Locale Conventions		Audience Appropriateness		Overall
		MisPunc	MisCap	Gram	Space	LineBrk	CharLim	LineLim	LocErr	LangDet	Profan	Formal	
en→zh	Gemma 3 4B	1.05	**0.00**	0.58	0.72	1.18	0.84	0.35	0.18	0.12	0.38	0.31	1.46
	DeepSeek-V3.2	0.61	**0.00**	0.19	0.34	0.82	0.52	0.14	**0.04**	**0.01**	**0.15**	**0.10**	0.93
	SMART (Gemma 3 4B)	0.15	0.03	0.08	0.17	0.03	0.03	0.03	0.08	0.06	0.24	0.13	0.67
	SMART (DeepSeek-V3.2)	0.11	**0.01**	**0.06**	0.15	**0.01**	**0.01**	**0.01**	0.06	0.04	0.17	**0.10**	0.53
	SMART (Judge: GPT-5.5)	**0.09**	**0.01**	**0.05**	**0.12**	**0.00**	**0.00**	**0.01**	**0.05**	**0.03**	**0.16**	**0.08**	**0.47**
	SMART (Claude 4.6)	**0.10**	**0.00**	**0.05**	**0.13**	**0.00**	**0.00**	**0.00**	**0.05**	**0.03**	**0.16**	**0.08**	**0.48**
en→de	Gemma 3 4B	1.24	0.65	0.86	0.68	2.45	1.98	**0.88**	0.68	0.28	0.64	0.52	2.21
	DeepSeek-V3.2	0.70	0.27	0.41	0.26	2.02	**1.61**	**0.52**	0.38	0.05	**0.36**	**0.25**	1.50
	SMART (Gemma 3 4B)	0.44	0.18	0.44	0.25	2.51	2.33	1.61	0.07	0.05	0.69	0.50	1.23
	SMART (DeepSeek-V3.2)	0.34	**0.16**	**0.32**	0.21	2.23	1.78	1.32	**0.05**	**0.03**	0.51	0.39	**0.99**
	SMART (Judge: GPT-5.5)	**0.27**	**0.14**	**0.27**	**0.20**	**1.80**	**1.59**	1.13	**0.04**	**0.02**	**0.45**	**0.33**	**0.87**
	SMART (Claude 4.6)	**0.28**	**0.14**	**0.27**	**0.19**	**1.85**	1.64	1.16	**0.04**	**0.02**	0.48	0.34	**0.87**
en→ko	Gemma 3 4B	1.02	**0.00**	0.82	0.71	1.14	0.82	0.45	0.28	0.15	1.15	0.92	1.82
	DeepSeek-V3.2	0.55	**0.00**	0.38	0.30	0.74	0.49	0.18	**0.06**	**0.02**	**0.76**	**0.56**	1.17
	SMART (Gemma 3 4B)	0.39	0.03	0.28	0.28	0.24	0.11	0.08	0.08	0.06	1.53	1.04	1.26
	SMART (DeepSeek-V3.2)	0.28	**0.01**	0.25	0.22	**0.20**	**0.09**	**0.06**	**0.06**	0.04	1.11	0.78	1.03
	SMART (Judge: GPT-5.5)	**0.27**	**0.00**	**0.23**	**0.21**	**0.17**	**0.08**	**0.05**	**0.05**	**0.03**	**0.88**	**0.60**	**0.91**
	SMART (Claude 4.6)	**0.26**	**0.00**	**0.22**	**0.20**	**0.17**	**0.08**	**0.05**	**0.05**	**0.03**	0.93	0.65	**0.92**
en→it	Gemma 3 4B	1.38	0.78	1.05	0.82	**2.38**	**1.88**	**0.82**	0.62	0.25	0.68	0.56	2.34
	DeepSeek-V3.2	0.86	0.39	0.58	0.36	**1.91**	**1.48**	**0.48**	0.34	**0.06**	**0.34**	**0.23**	1.62
	SMART (Gemma 3 4B)	0.52	0.24	0.43	0.39	3.70	3.82	2.82	0.27	0.09	0.58	0.36	1.56
	SMART (DeepSeek-V3.2)	0.44	0.20	0.36	0.30	2.92	2.90	2.28	**0.20**	**0.07**	0.48	0.30	**1.30**
	SMART (Judge: GPT-5.5)	**0.40**	**0.16**	**0.35**	**0.26**	2.94	2.44	1.87	**0.16**	**0.06**	0.46	**0.24**	**1.13**
	SMART (Claude 4.6)	**0.37**	**0.17**	**0.33**	**0.25**	2.76	2.44	1.91	**0.16**	**0.06**	**0.43**	0.27	**1.13**
en→es-ES	Gemma 3 4B	1.42	0.88	1.12	0.86	2.32	**1.82**	**0.78**	0.58	0.22	0.71	0.58	2.36
	DeepSeek-V3.2	0.91	0.49	0.64	0.41	**1.85**	**1.42**	**0.44**	0.26	**0.05**	**0.36**	**0.25**	1.62

Continued on next page

Table 16 (continued)

Dir.	Method / Setting	Linguistic Conventions				Technical			Locale Conventions		Audience Appropriateness		Overall
		MisPunc	MisCap	Gram	Space	LineBrk	CharLim	LineLim	LocErr	LangDet	Profan	Formal	
	SMART (Gemma 3 4B)	0.91	0.37	0.83	0.62	3.09	2.97	2.34	0.15	0.07	0.71	0.57	1.71
	SMART (DeepSeek-V3.2)	0.74	0.29	0.61	0.45	2.66	2.31	1.77	0.11	**0.05**	0.59	0.42	1.36
	SMART (Judge: GPT-5.5)	**0.64**	**0.28**	**0.47**	**0.37**	**2.27**	2.02	1.44	**0.09**	**0.04**	**0.51**	0.38	**1.17**
	SMART (Claude 4.6)	**0.66**	**0.26**	**0.52**	**0.40**	2.31	1.99	1.49	**0.10**	**0.04**	**0.51**	**0.35**	**1.18**
en→fr	Gemma 3 4B	1.40	0.82	1.08	0.84	**2.48**	**1.95**	**0.85**	0.48	0.24	0.74	0.60	2.31
	DeepSeek-V3.2	0.90	0.40	0.60	**0.35**	**2.01**	**1.55**	**0.52**	0.16	**0.03**	**0.40**	**0.27**	1.60
	SMART (Gemma 3 4B)	0.80	0.23	0.63	0.45	4.77	3.89	2.76	0.11	0.05	0.81	0.52	1.74
	SMART (DeepSeek-V3.2)	0.61	0.21	0.46	**0.35**	3.69	3.02	2.47	**0.09**	**0.03**	0.66	0.42	**1.40**
	SMART (Judge: GPT-5.5)	**0.51**	**0.19**	**0.41**	**0.35**	2.96	2.60	2.11	**0.09**	**0.02**	0.58	**0.35**	**1.19**
	SMART (Claude 4.6)	**0.52**	**0.18**	**0.42**	**0.32**	3.01	2.65	2.08	**0.08**	**0.02**	**0.53**	**0.35**	**1.19**
en→pt-PT	Gemma 3 4B	2.36	0.91	1.14	1.43	7.11	6.92	4.75	0.38	0.14	1.32	1.07	3.72
	DeepSeek-V3.2	1.66	0.68	0.91	0.96	5.43	5.24	3.38	0.25	0.11	1.01	0.82	2.73
	SMART (Gemma 3 4B)	1.11	0.41	0.69	0.71	3.54	3.81	2.25	0.18	0.07	1.00	0.55	1.89
	SMART (DeepSeek-V3.2)	0.83	0.34	**0.58**	0.54	2.66	2.88	**1.61**	**0.13**	**0.05**	0.72	0.46	1.50
	SMART (Judge: GPT-5.5)	**0.75**	**0.32**	**0.59**	**0.43**	**2.39**	**2.27**	**1.62**	**0.11**	**0.04**	**0.57**	**0.36**	**1.30**
	SMART (Claude 4.6)	**0.82**	**0.33**	**0.58**	**0.46**	**2.60**	**2.47**	**1.61**	**0.13**	**0.04**	**0.64**	**0.41**	**1.39**
en→pt-BR	Gemma 3 4B	1.86	0.52	1.26	0.83	4.47	5.61	3.70	0.21	0.11	0.98	0.80	2.56
	DeepSeek-V3.2	1.25	0.41	0.90	0.68	3.25	4.05	2.71	0.15	0.10	0.81	0.56	1.91
	SMART (Gemma 3 4B)	0.84	0.32	0.65	0.45	2.70	2.15	1.82	0.12	0.06	0.66	0.43	1.42
	SMART (DeepSeek-V3.2)	0.61	0.27	0.56	0.40	2.27	1.95	1.32	0.10	**0.04**	0.51	0.33	1.14
	SMART (Judge: GPT-5.5)	**0.52**	**0.22**	**0.48**	**0.36**	**1.82**	**1.92**	**1.18**	**0.07**	**0.03**	**0.39**	**0.30**	**0.97**
	SMART (Claude 4.6)	**0.60**	**0.23**	**0.50**	**0.37**	**2.00**	**1.90**	**1.26**	**0.09**	**0.03**	**0.45**	**0.31**	**1.04**
en→no	Gemma 3 4B	1.86	0.69	1.53	0.99	4.99	4.38	3.28	0.27	0.12	1.25	0.89	2.99
	DeepSeek-V3.2	1.45	0.52	1.11	0.70	3.61	3.49	2.32	0.21	0.11	0.84	0.68	2.18
	SMART (Gemma 3 4B)	0.88	0.35	0.68	0.59	2.86	2.44	1.84	0.15	0.07	0.71	0.51	1.60
	SMART (DeepSeek-V3.2)	0.73	**0.29**	**0.57**	0.46	2.27	**1.92**	1.53	**0.11**	**0.05**	0.52	0.40	1.29
	SMART (Judge: GPT-5.5)	**0.69**	**0.26**	**0.51**	**0.37**	**1.97**	**1.92**	**1.34**	**0.10**	**0.04**	**0.47**	**0.36**	**1.10**
	SMART (Claude 4.6)	**0.70**	**0.26**	**0.57**	**0.42**	**2.15**	**1.89**	**1.36**	**0.10**	**0.04**	**0.46**	**0.37**	**1.19**
en→da	Gemma 3 4B	2.29	0.77	1.47	0.97	6.96	4.70	4.01	0.25	0.12	1.41	0.93	3.30
	DeepSeek-V3.2	1.54	0.54	1.07	0.69	4.75	3.45	2.98	0.18	0.11	1.01	0.66	2.34
	SMART (Gemma 3 4B)	0.95	0.39	0.79	0.55	3.09	3.10	2.29	0.14	0.07	0.77	0.50	1.79
	SMART (DeepSeek-V3.2)	0.81	0.28	0.59	0.42	2.50	**2.28**	1.77	**0.10**	**0.05**	0.55	0.40	1.39
	SMART (Judge: GPT-5.5)	**0.73**	**0.26**	**0.49**	**0.33**	**2.34**	**1.99**	**1.68**	**0.09**	**0.04**	**0.45**	**0.33**	**1.19**
	SMART (Claude 4.6)	**0.74**	**0.27**	**0.54**	**0.39**	**2.32**	**1.99**	**1.64**	**0.10**	**0.04**	**0.52**	**0.39**	**1.26**
en→nl	Gemma 3 4B	1.93	0.70	1.49	1.16	5.79	6.82	4.40	0.32	0.12	1.70	1.23	3.41
	DeepSeek-V3.2	1.40	0.52	1.04	0.88	4.71	5.00	3.22	0.21	0.11	1.14	0.88	2.57
	SMART (Gemma 3 4B)	0.90	0.35	0.78	0.54	3.23	2.77	1.77	**0.13**	0.07	0.70	0.54	1.80
	SMART (DeepSeek-V3.2)	**0.75**	**0.31**	**0.59**	0.48	**2.43**	2.49	**1.59**	**0.11**	**0.05**	**0.54**	0.47	1.48
	SMART (Judge: GPT-5.5)	**0.66**	**0.29**	**0.55**	**0.43**	**2.30**	**2.15**	**1.44**	**0.11**	**0.04**	**0.54**	**0.38**	**1.29**
	SMART (Claude 4.6)	**0.75**	**0.29**	**0.55**	**0.47**	**2.43**	**2.33**	1.68	**0.11**	**0.04**	**0.53**	**0.43**	**1.38**
en→ro	Gemma 3 4B	1.62	0.66	1.64	0.94	5.74	3.76	3.44	0.23	0.11	1.04	0.80	2.96
	DeepSeek-V3.2	1.17	0.53	1.16	0.63	4.43	3.06	2.48	0.18	0.10	0.81	0.60	2.17
	SMART (Gemma 3 4B)	0.70	0.36	0.79	0.58	2.84	2.50	1.70	0.13	0.06	0.65	0.32	1.45
	SMART (DeepSeek-V3.2)	**0.61**	0.28	0.58	0.42	2.32	1.85	1.35	0.10	**0.04**	0.55	**0.28**	1.14
	SMART (Judge: GPT-5.5)	**0.57**	**0.26**	**0.48**	**0.31**	**1.96**	**1.68**	**1.18**	**0.08**	**0.03**	**0.48**	**0.29**	**1.01**
	SMART (Claude 4.6)	**0.57**	**0.25**	**0.51**	**0.36**	**2.14**	**1.63**	**1.27**	**0.09**	**0.03**	**0.47**	**0.28**	**1.07**
en→sv	Gemma 3 4B	1.40	0.44	1.08	0.87	6.18	4.54	3.39	0.21	0.11	1.00	0.87	2.57
	DeepSeek-V3.2	0.95	0.37	0.79	0.63	4.22	3.53	2.50	0.17	0.10	0.77	0.58	1.91
	SMART (Gemma 3 4B)	0.65	0.27	0.68	0.41	3.04	2.45	1.81	0.13	0.06	0.52	0.47	1.43
	SMART (DeepSeek-V3.2)	0.57	0.22	0.52	0.36	2.21	1.86	1.41	**0.11**	**0.04**	0.46	0.37	1.14
	SMART (Judge: GPT-5.5)	**0.50**	**0.18**	**0.42**	**0.33**	**1.82**	**1.83**	**1.35**	**0.09**	**0.03**	**0.35**	**0.33**	**0.95**
	SMART (Claude 4.6)	**0.52**	**0.20**	**0.44**	**0.35**	**1.98**	**1.76**	**1.38**	**0.09**	**0.03**	**0.41**	**0.32**	**1.04**
en→tr	Gemma 3 4B	1.63	0.58	1.34	0.92	5.51	5.80	3.73	0.26	0.11	1.37	0.79	3.04
	DeepSeek-V3.2	1.15	0.41	0.94	0.61	4.16	3.96	2.97	0.21	0.10	1.01	0.63	2.26
	SMART (Gemma 3 4B)	0.89	0.33	0.66	0.48	3.31	2.42	2.31	0.17	0.07	0.81	0.56	1.66
	SMART (DeepSeek-V3.2)	0.74	0.29	0.52	0.42	2.51	**2.02**	1.71	**0.13**	**0.05**	0.60	0.47	1.35
	SMART (Judge: GPT-5.5)	**0.67**	**0.22**	**0.50**	**0.34**	**2.32**	**1.87**	**1.50**	**0.11**	**0.04**	**0.45**	**0.35**	**1.17**
	SMART (Claude 4.6)	**0.71**	**0.25**	**0.51**	**0.38**	**2.27**	2.04	**1.47**	**0.11**	**0.04**	**0.52**	**0.40**	**1.25**
en→es-MX	Gemma 3 4B	1.68	0.74	1.37	1.22	6.58	5.71	3.91	0.28	0.14	1.07	0.79	3.43
	DeepSeek-V3.2	1.31	0.50	1.03	0.87	5.19	4.27	2.92	0.20	0.11	0.88	0.64	2.51
	SMART (Gemma 3 4B)	0.98	0.39	0.71	0.61	3.45	3.18	2.40	0.14	0.08	0.66	0.49	1.72
	SMART (DeepSeek-V3.2)	0.78	0.33	0.64	0.47	2.88	2.48	1.72	**0.10**	**0.06**	0.55	0.41	1.39
	SMART (Judge: GPT-5.5)	**0.67**	**0.30**	**0.59**	**0.42**	**2.49**	**2.06**	**1.49**	**0.09**	**0.05**	**0.50**	**0.34**	**1.23**
	SMART (Claude 4.6)	**0.76**	**0.29**	**0.63**	**0.45**	**2.47**	**2.21**	**1.59**	**0.09**	**0.05**	**0.52**	**0.36**	**1.31**

Abbreviations. Linguistic Conventions (MisPunc = Mispunctuation; MisCap = Miscapitalization; Gram = Grammar; Space = Spacing Error). **Technical** (LineBrk = Incorrect Line Breaking; CharLim = Exceeding Characters per Line; LineLim = Exceeding Lines per Box). **Locale Conventions** (LocErr = Localization Error; LangDet = Language Detection Error). **Audience Appropriateness** (Profan = Profanity; Formal = Formality Error).

Table 17: Fine-grained backbone and judge analysis on Subtitle Arena for translation from 15 source locales into English: Terminology, Accuracy, and Fluency. All values are penalties (lower is better). Best and second-best values within each language direction and metric are highlighted with blue-gray and warm-beige backgrounds, respectively. These results provide the full breakdown corresponding to Table 5.

Dir.	Method / Setting	Terminology		Accuracy			Fluency			Overall
		NameInc	TermInc	MisTrans	UndTrans	OvrTrans	Coher	Natur	Vivid	
zh→en	Gemma 3 4B	0.58	0.45	3.91	2.89	2.51	2.91	2.24	1.34	1.64
	DeepSeek-V3.2	0.36	0.26	2.74	1.86	1.50	2.11	1.39	0.78	1.07
	SMART (Gemma 3 4B)	0.25	0.13	1.25	1.09	1.02	1.44	1.04	0.69	0.62
	SMART (DeepSeek-V3.2)	**0.18**	0.11	1.05	0.87	0.85	1.17	0.94	0.61	**0.51**
	SMART (Judge: GPT-5.5)	**0.15**	**0.09**	**0.90**	**0.72**	**0.67**	**1.04**	**0.84**	**0.57**	**0.44**
	SMART (Claude 4.6)	**0.15**	**0.10**	**0.95**	**0.74**	**0.71**	**1.02**	**0.78**	**0.56**	**0.44**
de→en	Gemma 3 4B	0.53	0.39	3.58	2.68	2.32	2.60	1.99	1.18	1.49
	DeepSeek-V3.2	0.35	0.22	2.50	1.72	1.41	1.88	1.26	0.71	1.02
	SMART (Gemma 3 4B)	0.19	0.12	1.33	0.88	1.08	1.39	0.99	0.71	0.60
	SMART (DeepSeek-V3.2)	0.17	**0.10**	1.06	0.78	0.80	1.11	0.77	0.53	**0.47**
	SMART (Judge: GPT-5.5)	**0.13**	**0.09**	**0.94**	**0.73**	**0.67**	**0.89**	**0.65**	**0.49**	**0.41**
	SMART (Claude 4.6)	**0.14**	**0.09**	**0.89**	**0.68**	**0.65**	**0.93**	**0.70**	**0.50**	**0.41**
ko→en	Gemma 3 4B	0.61	0.48	4.11	3.09	2.71	3.08	2.39	1.41	1.74
	DeepSeek-V3.2	0.38	0.28	2.88	1.99	1.61	2.21	1.49	0.82	1.13
	SMART (Gemma 3 4B)	0.28	0.15	1.31	1.15	1.03	1.71	1.13	0.85	0.68
	SMART (DeepSeek-V3.2)	0.21	0.12	1.18	0.90	0.85	1.31	0.96	0.68	**0.55**
	SMART (Judge: GPT-5.5)	**0.16**	**0.11**	**0.96**	**0.80**	**0.77**	**1.12**	**0.78**	**0.62**	**0.47**
	SMART (Claude 4.6)	**0.17**	**0.10**	**1.01**	**0.78**	**0.74**	**1.06**	**0.81**	**0.59**	**0.47**
it→en	Gemma 3 4B	0.49	0.35	3.41	2.52	2.18	2.44	1.86	1.11	1.39
	DeepSeek-V3.2	0.31	0.20	2.34	1.59	1.32	1.76	1.18	0.67	0.91
	SMART (Gemma 3 4B)	0.19	0.10	1.13	0.96	0.94	1.40	0.89	0.60	0.55
	SMART (DeepSeek-V3.2)	0.16	**0.08**	1.00	0.73	0.73	1.04	0.70	0.54	**0.44**
	SMART (Judge: GPT-5.5)	**0.13**	**0.07**	**0.90**	**0.59**	**0.59**	**0.91**	**0.60**	**0.49**	**0.38**
	SMART (Claude 4.6)	**0.14**	**0.07**	**0.84**	**0.62**	**0.60**	**0.86**	**0.64**	**0.46**	**0.38**
es-ES→en	Gemma 3 4B	0.51	0.36	3.31	2.42	2.11	2.36	1.79	1.07	1.35
	DeepSeek-V3.2	0.32	0.21	2.28	1.52	1.26	1.70	1.12	0.64	0.87
	SMART (Gemma 3 4B)	0.23	0.10	1.18	0.74	0.98	1.05	0.87	0.57	0.52
	SMART (DeepSeek-V3.2)	**0.17**	0.08	0.90	0.65	0.71	0.90	0.68	0.52	0.41
	SMART (Judge: GPT-5.5)	**0.14**	**0.06**	**0.77**	**0.57**	**0.54**	**0.77**	**0.58**	**0.40**	**0.34**
	SMART (Claude 4.6)	**0.14**	**0.07**	**0.80**	**0.58**	**0.57**	**0.82**	**0.61**	**0.44**	**0.36**
fr→en	Gemma 3 4B	0.52	0.38	3.51	2.59	2.24	2.51	1.92	1.14	1.44
	DeepSeek-V3.2	0.33	0.22	2.41	1.64	1.36	1.81	1.22	0.68	0.94
	SMART (Gemma 3 4B)	0.24	0.12	1.31	0.95	0.90	1.29	0.95	0.72	0.58
	SMART (DeepSeek-V3.2)	**0.19**	**0.09**	1.05	0.73	0.66	0.96	0.78	0.57	**0.45**
	SMART (Judge: GPT-5.5)	**0.15**	**0.08**	**0.90**	**0.65**	**0.65**	**0.84**	**0.71**	**0.47**	**0.39**
	SMART (Claude 4.6)	**0.15**	**0.08**	**0.86**	**0.64**	**0.62**	**0.89**	**0.67**	**0.48**	**0.39**
pt-PT→en	Gemma 3 4B	0.57	0.23	2.68	1.39	1.72	2.58	1.83	1.26	1.09
	DeepSeek-V3.2	0.38	0.18	2.00	1.09	1.21	1.95	1.43	0.95	0.82
	SMART (Gemma 3 4B)	0.21	0.15	1.52	0.84	0.87	1.65	1.08	0.86	0.63
	SMART (DeepSeek-V3.2)	**0.16**	**0.11**	1.18	0.72	0.75	1.27	0.79	0.62	0.50
	SMART (Judge: GPT-5.5)	**0.14**	**0.08**	**1.03**	**0.63**	**0.67**	**1.17**	**0.71**	**0.53**	**0.44**
	SMART (Claude 4.6)	0.18	**0.08**	**0.97**	**0.62**	**0.61**	**0.92**	**0.69**	**0.49**	**0.41**
pt-BR→en	Gemma 3 4B	0.33	0.15	1.83	0.98	1.19	1.73	1.26	0.82	0.75
	DeepSeek-V3.2	0.25	0.12	1.33	0.72	**0.81**	1.33	0.90	0.68	0.55
	SMART (Gemma 3 4B)	0.19	0.14	1.52	1.12	1.02	1.39	1.13	0.91	0.66
	SMART (DeepSeek-V3.2)	0.17	0.10	1.13	0.86	0.83	1.09	0.91	0.68	0.51
	SMART (Judge: GPT-5.5)	**0.15**	**0.09**	**0.91**	**0.70**	**0.81**	**0.86**	**0.80**	**0.63**	**0.44**
	SMART (Claude 4.6)	**0.13**	**0.06**	**0.71**	**0.44**	**0.49**	**0.74**	**0.51**	**0.35**	**0.31**
no→en	Gemma 3 4B	0.30	0.20	1.71	1.44	1.61	1.99	1.25	1.13	0.86
	DeepSeek-V3.2	0.24	0.14	1.20	1.00	1.11	1.46	1.04	0.78	0.62
	SMART (Gemma 3 4B)	0.17	0.15	1.25	1.02	1.10	1.46	0.89	0.61	0.60
	SMART (DeepSeek-V3.2)	**0.15**	0.12	0.94	0.86	0.85	1.19	**0.76**	0.55	0.48
	SMART (Judge: GPT-5.5)	**0.13**	**0.11**	**0.81**	**0.68**	**0.74**	**1.02**	**0.76**	**0.46**	**0.42**
	SMART (Claude 4.6)	**0.13**	**0.07**	**0.68**	**0.50**	**0.53**	**0.75**	**0.54**	**0.39**	**0.32**
da→en	Gemma 3 4B	0.29	0.12	1.68	1.32	1.10	1.70	1.15	1.07	0.76
	DeepSeek-V3.2	0.21	**0.11**	1.16	1.00	0.91	1.34	0.90	0.74	0.57
	SMART (Gemma 3 4B)	0.22	0.13	1.56	1.00	1.11	1.49	1.06	0.88	0.66
	SMART (DeepSeek-V3.2)	0.18	**0.11**	1.21	0.78	0.95	1.24	0.91	0.72	0.54
	SMART (Judge: GPT-5.5)	**0.16**	**0.11**	**1.01**	**0.69**	**0.77**	**1.02**	**0.82**	**0.58**	**0.46**
	SMART (Claude 4.6)	**0.11**	**0.05**	**0.64**	**0.50**	**0.48**	**0.60**	**0.46**	**0.36**	**0.29**
nl→en	Gemma 3 4B	0.38	0.17	1.66	1.50	1.15	1.80	1.36	1.07	0.81
	DeepSeek-V3.2	0.26	0.14	1.33	1.06	**0.85**	1.32	**0.93**	0.77	0.60
	SMART (Gemma 3 4B)	0.30	0.17	1.51	1.03	1.33	1.51	1.30	0.98	0.73
	SMART (DeepSeek-V3.2)	0.22	0.12	1.28	**0.86**	1.00	1.33	1.01	0.84	0.59
	SMART (Judge: GPT-5.5)	**0.19**	**0.11**	**1.18**	**0.86**	**0.85**	**1.08**	0.94	**0.68**	**0.52**

Continued on next page

Table 17 (continued)

Dir.	Method / Setting	Terminology		Accuracy			Fluency			Overall
		NameInc	TermInc	MisTrans	UndTrans	OvrTrans	Coher	Natur	Vivid	
	SMART (Claude 4.6)	**0.12**	**0.06**	**0.70**	**0.52**	**0.47**	**0.71**	**0.52**	**0.36**	**0.31**
ro→en	Gemma 3 4B	0.36	0.17	2.11	1.41	1.22	2.27	1.22	1.04	0.87
	DeepSeek-V3.2	0.28	**0.13**	1.65	1.14	1.01	1.55	**0.96**	0.81	0.68
	SMART (Gemma 3 4B)	0.25	0.19	1.64	1.31	1.05	1.93	1.37	0.88	0.76
	SMART (DeepSeek-V3.2)	0.20	0.14	1.48	1.04	0.92	1.64	1.17	0.80	0.65
	SMART (Judge: GPT-5.5)	**0.17**	0.14	**1.24**	**0.84**	**0.80**	**1.32**	0.99	**0.66**	**0.55**
	SMART (Claude 4.6)	**0.14**	**0.07**	**0.80**	**0.58**	**0.54**	**0.77**	**0.54**	**0.39**	**0.34**
sv→en	Gemma 3 4B	0.41	0.18	2.18	1.29	1.62	2.24	1.83	1.36	0.98
	DeepSeek-V3.2	0.33	0.15	1.80	0.96	1.31	1.59	1.38	1.02	0.76
	SMART (Gemma 3 4B)	0.17	0.13	1.30	0.81	0.85	1.14	0.81	0.61	0.53
	SMART (DeepSeek-V3.2)	**0.15**	0.11	1.05	0.71	0.69	**0.92**	**0.65**	0.53	0.44
	SMART (Judge: GPT-5.5)	**0.13**	**0.09**	**0.81**	**0.66**	**0.62**	**0.88**	**0.55**	**0.44**	**0.38**
	SMART (Claude 4.6)	0.17	**0.08**	**0.92**	**0.59**	**0.61**	**0.88**	0.66	**0.52**	**0.39**
tr→en	Gemma 3 4B	0.37	0.20	2.17	1.76	1.70	2.03	1.60	1.52	1.02
	DeepSeek-V3.2	0.29	0.14	1.52	1.22	1.20	1.63	1.21	1.10	0.74
	SMART (Gemma 3 4B)	0.19	0.15	1.60	1.19	1.05	1.68	1.18	0.80	0.70
	SMART (DeepSeek-V3.2)	**0.17**	0.12	1.21	0.94	**0.86**	1.22	1.06	**0.60**	0.55
	SMART (Judge: GPT-5.5)	**0.14**	**0.10**	**1.01**	**0.89**	0.87	**1.05**	**0.90**	**0.60**	**0.49**
	SMART (Claude 4.6)	**0.14**	**0.07**	**0.83**	**0.66**	**0.58**	**0.86**	**0.64**	**0.51**	**0.38**
es-MX→en	Gemma 3 4B	0.35	0.18	2.44	1.45	1.22	1.61	1.46	1.22	0.90
	DeepSeek-V3.2	0.27	0.14	1.85	1.01	0.96	1.32	1.01	0.85	0.68
	SMART (Gemma 3 4B)	0.24	0.13	1.42	1.17	1.14	1.44	1.03	0.83	0.67
	SMART (DeepSeek-V3.2)	0.18	0.11	1.14	0.90	0.85	1.29	0.87	0.64	0.53
	SMART (Judge: GPT-5.5)	**0.15**	**0.10**	**1.05**	**0.70**	**0.77**	**1.02**	**0.72**	**0.58**	**0.45**
	SMART (Claude 4.6)	**0.13**	**0.07**	**0.88**	**0.60**	**0.55**	**0.79**	**0.61**	**0.43**	**0.36**

Abbreviations. Terminology (NameInc = Name Inconsistency; TermInc = Term Inconsistency). **Accuracy** (MisTrans = Mistranslation; UndTrans = Undertranslation; OvrTrans = Overtranslation). **Fluency** (Coher = Coherence; Natur = Naturalness; Vivid = Vividness).

Table 18: Fine-grained backbone and judge analysis on Subtitle Arena for translation from 15 source locales into English: Linguistic Conventions, Technical, Locale Conventions, and Audience Appropriateness. All values are penalties (lower is better). Best and second-best values within each language direction and metric are highlighted with blue-gray and warm-beige backgrounds, respectively. These results provide the full breakdown corresponding to Table 5.

Dir.	Method / Setting	Linguistic Conventions				Technical			Locale Conventions		Audience Appropriateness		Overall
		MisPunc	MisCap	Gram	Space	LineBrk	CharLim	LineLim	LocErr	LangDet	Profan	Formal	
zh→en	Gemma 3 4B	1.21	0.56	0.88	0.59	1.51	1.04	0.44	0.26	0.05	0.46	0.31	1.64
	DeepSeek-V3.2	0.64	0.32	0.46	0.24	1.41	0.99	0.24	0.15	**0.01**	0.25	0.10	1.07
	SMART (Gemma 3 4B)	0.29	0.05	0.21	0.07	0.21	0.05	0.08	0.04	0.06	0.18	0.06	0.62
	SMART (DeepSeek-V3.2)	0.22	**0.03**	0.17	**0.05**	**0.17**	**0.03**	**0.06**	**0.02**	0.04	**0.15**	**0.04**	**0.51**
	SMART (Judge: GPT-5.5)	**0.18**	**0.02**	**0.16**	**0.04**	**0.16**	**0.02**	**0.05**	**0.01**	**0.03**	**0.14**	**0.03**	**0.44**
	SMART (Claude 4.6)	**0.19**	**0.02**	**0.15**	**0.04**	**0.16**	**0.02**	**0.05**	**0.01**	**0.03**	**0.14**	**0.03**	**0.44**
de→en	Gemma 3 4B	1.01	0.46	0.74	0.46	1.41	0.92	0.38	0.32	0.05	0.39	0.24	1.49
	DeepSeek-V3.2	0.62	0.16	0.38	0.18	1.28	0.86	0.31	0.20	0.22	**0.08**	0.96	1.02
	SMART (Gemma 3 4B)	0.22	0.03	0.20	0.06	0.17	0.04	0.10	0.10	0.03	0.15	0.05	0.60
	SMART (DeepSeek-V3.2)	**0.20**	**0.01**	0.15	**0.04**	0.15	**0.02**	**0.08**	**0.08**	**0.01**	0.13	**0.03**	**0.47**
	SMART (Judge: GPT-5.5)	**0.17**	**0.00**	**0.14**	**0.03**	**0.13**	**0.01**	**0.07**	**0.08**	**0.00**	0.13	**0.02**	**0.41**
	SMART (Claude 4.6)	**0.17**	**0.00**	**0.13**	**0.03**	**0.14**	**0.01**	**0.07**	**0.07**	**0.00**	**0.12**	**0.02**	**0.41**
ko→en	Gemma 3 4B	1.26	0.62	0.94	0.64	1.56	1.09	0.46	0.29	0.05	0.50	0.34	1.74
	DeepSeek-V3.2	0.78	0.24	0.49	0.26	1.44	1.02	0.36	0.16	**0.00**	0.27	0.11	1.13
	SMART (Gemma 3 4B)	0.31	0.05	0.22	0.08	0.29	0.04	0.12	0.10	0.03	0.23	0.06	0.68
	SMART (DeepSeek-V3.2)	0.24	**0.03**	**0.17**	**0.06**	0.21	**0.02**	**0.10**	**0.08**	**0.01**	**0.20**	**0.04**	**0.55**
	SMART (Judge: GPT-5.5)	**0.22**	**0.02**	**0.16**	**0.05**	**0.20**	**0.01**	**0.09**	**0.08**	**0.00**	**0.16**	**0.03**	**0.47**
	SMART (Claude 4.6)	**0.21**	**0.02**	**0.16**	**0.05**	**0.19**	**0.01**	**0.09**	**0.07**	**0.00**	**0.16**	**0.03**	**0.47**
it→en	Gemma 3 4B	0.94	0.42	0.68	0.42	1.34	0.86	0.34	0.26	0.03	0.36	0.20	1.39
	DeepSeek-V3.2	0.57	0.14	0.37	0.14	1.22	0.80	0.28	0.18	**0.00**	0.21	0.07	0.91
	SMART (Gemma 3 4B)	0.23	0.03	0.16	0.05	0.22	0.03	0.10	0.09	0.03	0.17	0.03	0.55
	SMART (DeepSeek-V3.2)	**0.18**	**0.01**	**0.13**	**0.03**	**0.18**	**0.01**	**0.08**	**0.07**	**0.01**	0.14	**0.01**	**0.44**
	SMART (Judge: GPT-5.5)	**0.16**	**0.00**	**0.12**	**0.02**	**0.15**	**0.00**	**0.07**	**0.06**	**0.00**	**0.11**	**0.00**	**0.38**
	SMART (Claude 4.6)	**0.16**	**0.00**	**0.12**	**0.02**	**0.15**	**0.00**	**0.07**	**0.06**	**0.00**	**0.12**	**0.00**	**0.38**
es-ES→en	Gemma 3 4B	0.91	0.39	0.64	0.39	1.31	0.84	0.32	0.24	0.02	0.34	0.18	1.35
	DeepSeek-V3.2	0.54	0.12	0.34	0.12	1.20	0.78	0.27	0.16	**0.00**	0.20	0.06	0.87
	SMART (Gemma 3 4B)	0.18	0.03	0.18	0.04	0.23	0.03	0.09	0.09	0.03	0.16	0.03	0.52
	SMART (DeepSeek-V3.2)	0.16	**0.01**	**0.13**	**0.02**	0.17	**0.01**	**0.07**	**0.07**	**0.01**	0.13	**0.01**	0.41
	SMART (Judge: GPT-5.5)	**0.14**	**0.01**	**0.11**	**0.01**	**0.13**	**0.00**	**0.06**	**0.06**	**0.00**	**0.10**	**0.01**	**0.34**
	SMART (Claude 4.6)	**0.15**	**0.00**	**0.11**	**0.01**	**0.14**	**0.00**	**0.06**	**0.06**	**0.00**	**0.11**	**0.00**	**0.36**
fr→en	Gemma 3 4B	0.96	0.44	0.71	0.44	1.36	0.89	0.36	0.28	0.04	0.38	0.22	1.44
	DeepSeek-V3.2	0.58	0.14	0.38	0.14	1.24	0.82	0.30	0.19	**0.00**	0.22	0.07	0.94
	SMART (Gemma 3 4B)	0.27	0.03	0.16	0.04	0.20	0.04	0.11	0.10	0.03	0.20	0.03	0.58
	SMART (DeepSeek-V3.2)	0.20	**0.01**	**0.13**	**0.02**	**0.16**	**0.02**	**0.09**	**0.08**	**0.01**	0.15	**0.01**	**0.45**
	SMART (Judge: GPT-5.5)	**0.16**	**0.00**	**0.12**	**0.01**	**0.15**	**0.01**	**0.08**	**0.07**	**0.01**	**0.12**	**0.01**	**0.39**

Continued on next page

Table 18 (continued)

Dir.	Method / Setting	Linguistic Conventions				Technical			Locale Conventions		Audience Appropriateness		Overall
		MisPunc	MisCap	Gram	Space	LineBrk	CharLim	LineLim	LocErr	LangDet	Profan	Formal	
	SMART (Claude 4.6)	**0.17**	**0.00**	**0.12**	**0.01**	**0.15**	**0.01**	**0.08**	**0.07**	**0.00**	**0.13**	**0.00**	**0.39**
pt-PT→en	Gemma 3 4B	0.52	0.07	0.31	0.08	0.38	0.07	0.16	0.19	0.08	0.31	0.08	1.09
	DeepSeek-V3.2	0.37	0.06	0.25	0.07	0.27	0.06	0.13	0.13	0.07	0.24	0.07	0.82
	SMART (Gemma 3 4B)	0.33	0.05	0.21	0.07	0.25	0.05	0.08	0.04	0.06	0.19	0.06	0.63
	SMART (DeepSeek-V3.2)	0.26	0.03	0.17	0.05	0.21	0.03	**0.06**	**0.02**	0.04	**0.16**	0.04	0.50
	SMART (Judge: GPT-5.5)	**0.21**	**0.02**	**0.16**	**0.04**	**0.19**	**0.02**	**0.05**	**0.01**	**0.03**	**0.16**	**0.03**	**0.44**
	SMART (Claude 4.6)	**0.17**	**0.00**	**0.14**	**0.01**	**0.17**	**0.00**	0.07	0.07	**0.01**	**0.14**	**0.01**	**0.41**
pt-BR→en	Gemma 3 4B	0.34	0.07	0.19	0.08	0.31	0.07	0.15	0.12	0.08	0.26	0.08	0.75
	DeepSeek-V3.2	0.25	0.06	0.15	0.07	0.23	0.06	0.11	0.11	0.07	0.20	0.07	0.55
	SMART (Gemma 3 4B)	0.21	0.05	0.17	0.07	0.25	0.05	0.08	0.04	0.06	0.22	0.06	0.66
	SMART (DeepSeek-V3.2)	**0.17**	0.03	0.15	0.05	0.19	0.03	**0.06**	**0.02**	0.04	0.17	0.04	0.51
	SMART (Judge: GPT-5.5)	**0.17**	**0.02**	**0.13**	**0.04**	**0.15**	**0.02**	**0.05**	**0.01**	**0.03**	**0.15**	**0.03**	**0.44**
	SMART (Claude 4.6)	**0.13**	**0.00**	**0.09**	**0.01**	**0.12**	**0.00**	**0.05**	0.05	**0.01**	**0.10**	**0.01**	**0.31**
no→en	Gemma 3 4B	0.28	0.08	0.28	0.08	0.32	0.08	0.12	0.12	0.08	0.25	0.08	0.86
	DeepSeek-V3.2	0.23	0.07	0.19	0.07	0.23	0.07	0.11	0.11	0.07	0.20	0.07	0.62
	SMART (Gemma 3 4B)	0.22	0.05	0.21	0.07	0.19	0.05	0.08	0.04	0.06	0.18	0.06	0.60
	SMART (DeepSeek-V3.2)	**0.19**	0.03	0.17	0.05	0.16	0.03	**0.06**	**0.02**	0.04	0.15	0.04	0.48
	SMART (Judge: GPT-5.5)	**0.19**	**0.02**	**0.15**	**0.04**	**0.14**	**0.02**	**0.05**	**0.01**	**0.03**	**0.14**	**0.03**	**0.42**
	SMART (Claude 4.6)	**0.13**	**0.01**	**0.09**	**0.01**	**0.13**	**0.01**	**0.05**	0.05	**0.00**	**0.10**	**0.01**	**0.32**
da→en	Gemma 3 4B	0.30	0.08	0.20	0.08	0.25	0.08	0.16	0.16	0.08	0.22	0.07	0.76
	DeepSeek-V3.2	0.22	0.07	0.16	0.07	**0.18**	0.07	0.11	0.11	0.07	0.15	0.06	0.57
	SMART (Gemma 3 4B)	0.30	0.05	0.21	0.07	0.23	0.05	0.10	0.04	0.06	0.17	0.06	0.66
	SMART (DeepSeek-V3.2)	0.23	0.03	0.17	0.05	**0.18**	0.03	0.07	**0.02**	0.04	0.15	0.04	0.54
	SMART (Judge: GPT-5.5)	**0.18**	**0.02**	**0.14**	**0.04**	**0.18**	**0.02**	**0.06**	**0.01**	**0.03**	**0.14**	**0.03**	**0.46**
	SMART (Claude 4.6)	**0.12**	**0.01**	**0.09**	**0.01**	**0.11**	**0.01**	**0.05**	0.05	**0.00**	**0.09**	**0.00**	**0.29**
nl→en	Gemma 3 4B	0.34	0.08	0.22	0.08	0.28	0.08	0.12	0.14	0.07	0.19	0.08	0.81
	DeepSeek-V3.2	0.23	0.07	**0.18**	0.07	0.21	0.07	0.11	0.11	0.06	**0.15**	0.07	0.60
	SMART (Gemma 3 4B)	0.36	0.06	0.26	0.07	0.30	0.05	0.09	0.04	0.06	0.23	0.06	0.73
	SMART (DeepSeek-V3.2)	0.28	0.04	0.20	0.05	0.24	0.03	0.07	**0.02**	0.04	0.20	0.04	0.59
	SMART (Judge: GPT-5.5)	**0.21**	**0.03**	**0.18**	**0.04**	**0.19**	**0.02**	**0.06**	**0.01**	**0.03**	**0.15**	**0.03**	**0.52**
	SMART (Claude 4.6)	**0.12**	**0.01**	**0.09**	**0.01**	**0.12**	**0.01**	**0.05**	0.05	**0.00**	**0.09**	**0.01**	**0.31**
ro→en	Gemma 3 4B	0.33	0.07	0.22	0.08	0.31	0.07	0.19	0.14	0.08	0.25	0.07	0.87
	DeepSeek-V3.2	0.23	0.06	**0.15**	0.07	0.24	0.06	0.13	0.11	0.07	**0.17**	0.06	0.68
	SMART (Gemma 3 4B)	0.33	0.06	0.25	0.08	0.27	0.05	0.09	0.04	0.07	0.25	0.07	0.76
	SMART (DeepSeek-V3.2)	0.24	0.04	0.19	0.06	0.22	0.03	0.07	**0.02**	0.05	0.22	0.05	0.65
	SMART (Judge: GPT-5.5)	**0.21**	**0.03**	0.16	**0.05**	**0.19**	**0.02**	**0.06**	**0.01**	**0.04**	0.19	**0.04**	**0.55**
	SMART (Claude 4.6)	**0.13**	**0.00**	**0.09**	**0.01**	**0.14**	**0.00**	**0.05**	0.05	**0.01**	**0.10**	**0.00**	**0.34**
sv→en	Gemma 3 4B	0.45	0.08	0.27	0.08	0.32	0.07	0.16	0.17	0.08	0.25	0.08	0.98
	DeepSeek-V3.2	0.31	0.07	0.21	0.07	0.25	0.06	0.12	0.13	0.07	0.19	0.07	0.76
	SMART (Gemma 3 4B)	0.30	0.05	0.18	0.07	0.18	0.05	0.07	0.04	0.06	0.16	0.06	0.53
	SMART (DeepSeek-V3.2)	0.22	0.03	**0.14**	0.05	**0.16**	0.03	**0.05**	**0.02**	0.04	0.14	0.04	0.44
	SMART (Judge: GPT-5.5)	**0.17**	**0.02**	**0.13**	**0.04**	**0.12**	**0.02**	**0.04**	**0.01**	**0.03**	**0.13**	**0.03**	**0.38**
	SMART (Claude 4.6)	**0.16**	**0.01**	**0.13**	**0.01**	**0.16**	**0.00**	0.06	0.07	**0.01**	**0.11**	**0.01**	**0.39**
tr→en	Gemma 3 4B	0.37	0.08	0.31	0.08	0.31	0.08	0.14	0.21	0.08	0.28	0.08	1.02
	DeepSeek-V3.2	0.27	0.07	0.23	0.07	0.23	0.07	0.12	0.15	0.07	0.19	0.07	0.74
	SMART (Gemma 3 4B)	0.30	0.05	0.24	0.07	0.26	0.05	0.08	0.04	0.06	0.21	0.06	0.70
	SMART (DeepSeek-V3.2)	0.25	0.03	0.20	0.05	**0.20**	0.03	**0.06**	**0.02**	0.04	0.18	0.04	0.55
	SMART (Judge: GPT-5.5)	**0.19**	**0.02**	**0.16**	**0.04**	0.21	**0.02**	**0.05**	**0.01**	**0.03**	**0.14**	**0.03**	**0.49**
	SMART (Claude 4.6)	**0.15**	**0.01**	**0.13**	**0.01**	**0.14**	**0.01**	**0.06**	0.07	**0.01**	**0.11**	**0.01**	**0.38**
es-MX→en	Gemma 3 4B	0.33	0.07	0.24	0.08	0.35	0.08	0.14	0.21	0.08	0.34	0.08	0.90
	DeepSeek-V3.2	0.26	0.06	0.19	0.07	0.26	0.07	0.12	0.16	0.07	0.23	0.07	0.68
	SMART (Gemma 3 4B)	0.31	0.05	0.22	0.07	0.23	0.05	0.08	0.04	0.06	0.23	0.06	0.67
	SMART (DeepSeek-V3.2)	0.23	0.03	0.17	0.05	0.17	0.03	**0.06**	**0.02**	0.04	0.18	0.04	0.53
	SMART (Judge: GPT-5.5)	**0.21**	**0.02**	**0.13**	**0.04**	**0.15**	**0.02**	**0.05**	**0.01**	**0.03**	**0.16**	**0.03**	**0.45**
	SMART (Claude 4.6)	**0.15**	**0.00**	**0.12**	**0.01**	**0.14**	**0.00**	**0.06**	0.07	**0.01**	**0.13**	**0.00**	**0.36**

Abbreviations. Linguistic Conventions (MisPunc = Mispunctuation; MisCap = Miscapitalization; Gram = Grammar; Space = Spacing Error). **Technical** (LineBrk = Incorrect Line Breaking; CharLim = Exceeding Characters per Line; LineLim = Exceeding Lines per Box). **Locale Conventions** (LocErr = Localization Error; LangDet = Language Detection Error). **Audience Appropriateness** (Profan = Profanity; Formal = Formality Error).

F.4 Temporal Generalization to Newly Released TV Series

To test whether SMART's gains extend beyond the content represented in Subtitle Arena, we construct a separate evaluation set of 200 TV series released between 2025 and 2026. This set was collected independently and is disjoint from Subtitle Arena in title. Anthropic reports a May 2025 knowledge cutoff for Claude Sonnet 4.6 (Anthropic, 2026a); therefore, the collection includes titles released beyond the backbone's reported knowledge horizon. We use this evaluation to test transfer to newly released television content rather than differences in translation difficulty across historical eras.

Table 19 reports SubMQM results on the same new-series split for Online subtitles, TransAgent, and SMART. SMART achieves a lower Overall penalty than TransAgent in all six directions. Its average Overall penalty is 0.73, compared with 0.77 for TransAgent, with absolute improvements of 0.04, 0.07, 0.02, 0.03, 0.02, and 0.01 for en→zh, en→de, en→ko, en→it, en→es, and en→fr, respectively.

Table 19: SubMQM results on 200 newly collected TV series released in 2025–2026. All values are penalties (lower is better). The best and second-best values for each metric within each direction are highlighted in blue-gray and warm beige, respectively.

Direction	Method	Term.	Acc.	Flu.	Ling.	Tech.	Loc.	Aud.	Overall
en→zh	Online	0.28	3.52	2.95	1.42	0.12	0.85	**0.08**	1.87
	TransAgent	**0.17**	**0.88**	**0.79**	**0.04**	**0.05**	**0.04**	0.11	**0.48**
	SMART	**0.16**	**0.83**	**0.71**	**0.03**	**0.00**	**0.05**	**0.10**	**0.44**
en→de	Online	0.55	4.37	**2.63**	0.51	1.45	0.67	**0.29**	2.14
	TransAgent	**0.28**	**1.51**	**1.46**	**0.31**	**1.07**	**0.05**	**0.40**	**0.94**
	SMART	**0.27**	**1.24**	**1.46**	**0.26**	**1.27**	**0.03**	0.42	**0.87**
en→ko	Online	0.44	2.73	2.06	0.93	0.30	**0.15**	0.56	1.48
	TransAgent	**0.17**	**0.62**	**0.81**	**0.13**	**0.12**	**0.03**	**0.35**	**0.44**
	SMART	**0.15**	**0.59**	**0.74**	**0.12**	**0.09**	**0.03**	**0.42**	**0.42**
en→it	Online	**0.45**	4.21	3.15	**2.74**	**0.98**	0.79	**0.30**	2.33
	TransAgent	**0.20**	**1.43**	**1.35**	**0.16**	**1.01**	**0.13**	**0.07**	**0.81**
	SMART	**0.20**	**1.23**	**1.43**	**0.16**	1.05	**0.11**	**0.07**	**0.78**
en→es	Online	0.41	5.97	3.58	1.30	**0.98**	0.46	0.77	2.86
	TransAgent	**0.19**	**1.83**	**1.09**	**0.49**	0.99	**0.06**	**0.19**	**0.92**
	SMART	**0.20**	**1.61**	**1.26**	**0.53**	**0.94**	**0.05**	**0.21**	**0.90**
en→fr	Online	0.51	4.21	2.94	1.82	0.93	**0.66**	0.75	2.27
	TransAgent	**0.13**	**1.49**	**2.06**	**0.40**	**0.71**	**0.09**	**0.36**	**1.00**
	SMART	**0.12**	**1.47**	**2.05**	**0.41**	**0.69**	**0.09**	**0.34**	**0.99**

Abbreviations. Term. = Terminology; Acc. = Accuracy; Flu. = Fluency; Ling. = Linguistic Conventions; Tech. = Technical; Loc. = Locale Conventions; Aud. = Audience Appropriateness.

The largest difference appears in semantic fidelity: averaged across the six directions, the Accuracy penalty decreases from 1.29 with TransAgent to 1.16 with SMART. Terminology and Locale Conventions remain comparable, while Fluency, Technical, and Audience Appropriateness show smaller mixed differences. These results show that the performance pattern observed on Subtitle Arena persists on a temporally shifted set containing newly released television content.

F.5 Ablation Settings

We provide detailed definitions of the ablation variants used in Table 6. Unless otherwise specified, each variant changes only the indicated component while retaining the remaining SMART configuration, backbone models, prompts, tools, and evaluation protocol.

w/o Dynamic Router. We replace the dynamic routing policy with a fixed translation workflow. Instead of selecting translator roles and auxiliary tools according to the current subtitle context, all inputs follow the same predefined execution path.

w/o Self-Evolution. We disable SMART's test-time evolution stage. The routing policy and agent prompts remain at their initial configurations throughout evaluation rather than being updated using feedback accumulated from the series-level test-time-training subset.

w/o Memory. We remove the persistent series memory shared across subtitle sentences and episodes. Each translation is therefore produced without access to previously accumulated series-level information, while the remaining retrieval, routing, and refinement components are unchanged.

w/o Contextual Retrieval & Idiom Bank. We disable the retrieval mechanisms that provide contextually relevant information and previously accumulated idiomatic or expression-level knowledge to the translation agents. The agents still receive their ordinary subtitle context and can use the remaining components of SMART.

w/o Sliding Window. We remove the series-level sliding-window consistency stage. Translation and refinement are performed without the additional overlapping-window pass used to identify and correct inconsistencies across neighboring subtitle sentences.

w/o MoA. We replace the Mixture-of-Agents translator with a single translator, removing parallel specialist hypothesis generation and subsequent candidate selection. All other routing, memory, retrieval, and refinement components are retained.

w/ Combined Prompt. To distinguish the effect of multiple independently generated hypotheses from the effect of specialist instructions themselves, we concatenate the instructions used by the specialist translators into a single prompt and use one translator to produce the translation. This variant therefore exposes the model to the same high-level translation guidance while removing independent candidate generation and selection.

F.6 Computational Cost and Latency

Table 20 reports the computational overhead of SMART. An episode contains 611.7 subtitle sentences on average and requires approximately 4,970 model calls, 14.03M input tokens, 487.7K output tokens, and 3,501 tool calls. Under serial execution, the average end-to-end runtime is 25.94 minutes per episode.

These results make the quality-compute trade-off explicit. SMART achieves stronger translation quality through substantially more test-time computation than a single-call translator. At the same time, runtime remains relatively stable across the six evaluated languages, with all directions completing in approximately 22 to 27 minutes per episode. SMART is therefore most suitable for long-form translation settings where semantic accuracy, contextual consistency, and subtitle quality are more important than minimizing inference-time computation.

Table 20: Average computational cost per episode across six translation directions. Runtime is measured under serial execution.

Direction	sentences / Ep.	API Calls / Ep.	Input Tokens / Ep.	Output Tokens / Ep.	Tool Calls / Ep.	Time / Ep. (min)
en→zh	571.3	4545.9	12.83M	446.1K	2524.7	21.78
en→es	597.2	4881.1	13.78M	479.0K	3826.7	25.60
en→fr	625.4	5453.9	15.78M	535.2K	3275.3	27.34
en→de	625.4	5312.2	14.99M	521.3K	4151.4	26.97
en→it	625.4	4229.7	11.94M	415.1K	3025.7	26.75
en→ko	625.4	5394.9	15.23M	529.5K	4201.9	27.19
Average	611.7	4969.6	14.03M	487.7K	3500.9	25.94

F.7 Human Evaluation Details

To reduce annotation burden, we do not ask human annotators to reproduce the full 19-error-type SubMQM rubric. Instead, we consolidate the fine-grained criteria into four high-level dimensions that are easier to assess consistently over continuous subtitle passages. **Fidelity** measures whether the translation preserves the meaning, intent, and relevant information of the source without omission, addition, or semantic distortion. **Consistency** measures whether recurring names, terminology, character references, and discourse choices remain consistent across the passage. **Language Quality** covers fluency, naturalness, stylistic appropriateness, and linguistic conventions in the target language. **Subtitle Quality** captures viewer-facing presentation quality, including readability, conciseness, segmentation, and compatibility with subtitle display constraints.

Twenty annotators participate in the study. The evaluation contains 20 continuous passages comprising 267 aligned subtitle lines from 13 television series. Each passage is presented together with its corresponding video clip and English source subtitles. For every passage, annotators compare five anonymized candidate translations produced by Online, Claude Sonnet 4.6, GPT-5.5, TransAgent, and SMART. Candidate order is independently randomized for each annotation instance to reduce position bias.

For each of the four criteria, annotators rank the five candidate translations from best to worst. The highest-ranked candidate receives a score of 5, followed by 4, 3, 2, and 1 for the lowest-ranked candidate. Annotators additionally provide an **Overall Preference** ranking after considering the translation as a whole. We report the average ranking score across annotations and passages, such that higher values indicate stronger human preference.

Table 21: Fine-grained SubMQM results for multimodal expansion on six representative translation directions. We compare text-only SMART with ViDove, Hermes, and SMART augmented with video/audio context (**SMART + MM**). All values are penalties (lower is better). For clarity, only the best and second-best Overall scores within each translation direction are highlighted with blue-gray and warm-beige backgrounds, respectively.

Dir.	Method	Terminology		Accuracy			Fluency			Linguistic Conventions				Technical			Locale Conventions	Audience Appropriateness		Overall
		Name	Term	Mis.	Under	Over	Coh.	Nat.	Viv.	MisP.	MisC.	Gram.	Space	Brk.	CPL	Lines	Loc.	Lang.	Prof.	Form.	
en→zh	SMART	0.18	0.12	0.86	0.81	0.73	1.08	0.94	0.83	0.10	0.00	0.05	0.13	0.00	0.00	0.00	0.05	0.03	0.16	0.08	**0.48**
	ViDove	0.18	0.14	0.72	0.87	0.72	1.36	1.12	0.62	0.11	0.00	0.06	0.15	0.00	0.00	0.00	0.05	0.03	0.17	0.08	0.49
	Hermes	0.19	0.13	0.64	0.69	0.78	1.57	1.01	0.65	0.09	0.00	0.06	0.15	0.06	0.00	0.06	0.05	0.03	0.17	0.07	**0.48**
	SMART + MM	0.18	0.12	0.72	0.70	0.73	1.08	0.94	0.59	0.10	0.00	0.05	0.13	0.00	0.00	0.00	0.03	0.03	0.16	0.08	**0.44**
en→de	SMART	0.29	0.21	1.39	1.28	1.05	1.62	1.46	1.30	0.28	0.14	0.27	0.19	1.85	1.64	1.16	0.04	0.02	0.48	0.34	0.87
	ViDove	0.33	0.20	1.19	1.23	1.07	2.44	1.64	1.05	0.29	0.14	0.34	0.19	2.05	1.58	1.33	0.04	0.02	0.46	0.38	0.91
	Hermes	0.28	0.24	1.00	0.97	0.98	2.62	1.47	0.79	0.30	0.16	0.28	0.18	2.51	1.83	1.24	0.04	0.02	0.51	0.35	**0.85**
	SMART + MM	0.29	0.21	0.97	0.99	1.05	1.62	1.46	0.95	0.28	0.14	0.27	0.19	1.85	1.64	1.16	0.03	0.02	0.48	0.34	**0.78**
en→fr	SMART	0.20	0.10	1.82	1.58	1.43	2.41	2.13	1.97	0.52	0.18	0.42	0.32	3.01	2.65	2.08	0.08	0.02	0.53	0.35	**1.19**
	ViDove	0.20	0.10	1.37	1.55	1.62	3.38	2.61	1.63	0.58	0.18	0.50	0.34	3.35	2.99	2.11	0.08	0.02	0.55	0.39	1.25
	Hermes	0.19	0.10	1.31	1.30	1.51	4.32	2.16	1.28	0.61	0.18	0.44	0.32	3.98	2.46	2.24	0.08	0.02	0.53	0.40	1.22
	SMART + MM	0.20	0.10	1.27	1.31	1.43	2.41	2.13	1.38	0.52	0.18	0.42	0.32	3.01	2.65	2.08	0.05	0.02	0.53	0.35	**1.06**
zh→en	SMART	0.15	0.10	0.95	0.74	0.71	1.02	0.78	0.56	0.19	0.02	0.15	0.04	0.16	0.02	0.05	0.01	0.03	0.14	0.03	**0.44**
	ViDove	0.16	0.11	0.73	0.72	0.79	1.60	1.06	0.42	0.20	0.02	0.18	0.04	0.16	0.02	0.05	0.01	0.03	0.16	0.03	0.48
	Hermes	0.16	0.11	0.67	0.57	0.70	1.66	0.90	0.40	0.21	0.02	0.15	0.04	0.19	0.02	0.08	0.01	0.03	0.13	0.03	**0.44**
	SMART + MM	0.15	0.10	0.75	0.60	0.71	1.02	0.78	0.37	0.19	0.02	0.15	0.04	0.16	0.02	0.05	0.00	0.03	0.14	0.03	**0.40**
de→en	SMART	0.14	0.09	0.89	0.68	0.65	0.93	0.70	0.50	0.17	0.00	0.13	0.03	0.14	0.01	0.07	0.07	0.00	0.12	0.02	**0.41**
	ViDove	0.15	0.09	0.69	0.75	0.70	1.38	0.83	0.39	0.16	0.00	0.14	0.03	0.16	0.01	0.08	0.08	0.00	0.13	0.02	0.43
	Hermes	0.15	0.09	0.70	0.52	0.75	1.49	0.65	0.37	0.18	0.00	0.15	0.03	0.24	0.01	0.10	0.07	0.00	0.13	0.02	**0.41**
	SMART + MM	0.14	0.09	0.73	0.57	0.65	0.93	0.70	0.33	0.17	0.00	0.13	0.03	0.14	0.01	0.07	0.05	0.00	0.12	0.02	**0.37**
fr→en	SMART	0.15	0.08	0.86	0.64	0.62	0.89	0.67	0.48	0.17	0.00	0.12	0.01	0.15	0.01	0.08	0.07	0.00	0.13	0.00	**0.39**
	ViDove	0.16	0.08	0.72	0.71	0.69	1.17	0.86	0.37	0.18	0.00	0.13	0.01	0.17	0.01	0.08	0.07	0.00	0.14	0.00	0.42
	Hermes	0.14	0.09	0.61	0.48	0.65	1.59	0.65	0.33	0.16	0.00	0.13	0.01	0.24	0.01	0.13	0.07	0.00	0.15	0.00	**0.39**
	SMART + MM	0.15	0.08	0.59	0.47	0.62	0.89	0.67	0.26	0.17	0.00	0.12	0.01	0.15	0.01	0.08	0.04	0.00	0.13	0.00	**0.34**

Abbreviations. Terminology: Name = Name Inconsistency; Term = Term Inconsistency. **Accuracy**: Mis. = Mistranslation; Under = Undertranslation; Over = Overtranslation. **Fluency**: Coh. = Coherence; Nat. = Naturalness; Viv. = Vividness. **Linguistic Conventions**: MisP. = Mispunctuation; MisC. = Miscapitalization; Gram. = Grammar; Space = Spacing Error. **Technical**: Brk. = Incorrect Line Breaking; CPL = Exceeding Characters per Line; Lines = Exceeding Lines per Box. **Locale Conventions**: Loc. = Localization Error; Lang. = Language Detection Error. **Audience Appropriateness**: Prof. = Profanity; Form. = Formality Error.

The human study is intended as a complementary validation of viewer-facing translation quality rather than a replacement for SubMQM. Several fine-grained error categories, such as punctuation, capitalization, locale-specific formatting, language detection, and character-per-line violations, are localized or relatively sparse and are difficult to estimate reliably with a study of this scale. Human evaluation therefore focuses on broad perceptual quality, while the LLM-based SubMQM evaluation is retained for scalable and fine-grained diagnostic analysis. The annotation interface and representative evaluation examples are provided in Figure 3.

G Multimodal Expansion

SMART primarily operates on text-based SRT subtitles, but its tool-augmented design allows additional sources of contextual evidence to be incorporated without modifying the underlying translation workflow. To evaluate this extensibility, we introduce multimodal tools that expose information from the video and audio streams aligned with each subtitle segment.

Specifically, we use Qwen3-Omni (Xu et al., 2025b) as the multimodal backbone. Given the timestamp of a subtitle segment, the tools retrieve its corresponding video clip or audio span and return a textual description of the relevant audiovisual evidence. These tools are exposed to the MOA Translators together with the existing textual context and retrieval tools. Importantly, multimodal processing is invoked selectively: translators may query audiovisual evidence when the subtitle alone leaves an ambiguity unresolved, rather than processing the entire video indiscriminately. Typical cases include identifying the active speaker, resolving visually grounded pronouns or references, interpreting emotional delivery, distinguishing literal from sarcastic utterances, and inferring scene-dependent expressions.

We compare four systems on six representative translation directions: text-only SMART, ViDove (Lu et al., 2025), Hermes (Cui et al., 2026a), and our multimodal extension (**SMART + MM**). Table 21 reports the complete SubMQM breakdown.

Across all six directions, SMART + MM achieves the lowest overall SubMQM penalty. Averaged across directions, multimodal augmentation reduces the overall penalty from 0.63 for text-only

SMART to 0.57, corresponding to a relative reduction of approximately 9.5%. The largest improvements tend to occur in dimensions that depend strongly on contextual interpretation, particularly mistranslation, undertranslation, vividness, and locale-related errors. In contrast, dimensions governed largely by surface-form constraints, such as punctuation, capitalization, spacing, and subtitle line formatting, change little.

Figure 4 shows what the reduction looks like on a single line. In a forensic examination scene, the source line See the striation? leaves both the number and the referent of the mark unspecified. Text-only SMART must choose from the text alone and produces a generic plural, whereas the multimodal tools report that a single mark on a bone is under the magnifier, and the translation names it accordingly. The error is one the text-only system has no evidence to avoid, and it falls in the Accuracy dimension where the table shows the largest average gain.

The results also illustrate an advantage of exposing multimodal information as tools rather than making multimodal processing mandatory for every sentence. Most subtitle sentences can already be translated accurately from textual and long-range discourse context, while only a subset benefits materially from inspecting the underlying video or audio. SMART therefore retains its original translation workflow and invokes multimodal reasoning only when additional evidence is useful.

H Qualitative Analysis on Rendered Frames

The SubMQM tables report penalties and the case studies compare passages as text, but neither shows what the differences look like where a viewer actually meets them: on the screen, for the two or three seconds the line is up. We therefore take six lines from the English→Chinese passages of the human study (Appendix F.7) and composite each system's subtitle onto the video frame at the moment the line is spoken, at the position, size, and two-line budget a player would use. The English source is burned into each panel directly beneath the Chinese, at a smaller size, so that every comparison can be read on its own. Within a row all panels are the identical frame, so the target-language subtitle is the only variable. Figures 5 and 6 show the result, and Table 22 indexes the six lines.

The lines are drawn from Nikita S01E01, which contributes two of the study passages: one of ordinary dialogue (00:25:04–00:25:40) and one requiring external knowledge about the series (00:34:05–00:34:57). Five systems are shown: Online, single calls to Claude Sonnet 4.6 and Claude Opus 4.8, the agentic baseline TransAgent, and SMART. Claude Sonnet 4.6 is SMART's own backbone, so that column isolates the scaffold from the underlying model, and the TransAgent column separates SMART's margin over a fixed pipeline of agents from its margin over a single call. Two caveats apply. The panels are our own compositing rather than a screenshot of the annotation interface, in which candidate translations were presented as unlabeled text columns beside the clip rather than burned into the video. And six lines chosen to be legible in print illustrate the error types that the tables aggregate; they are not evidence about how often those errors occur.

Construction. The panels are built from the same material as the annotation interface: the clip files shown to annotators, the candidate texts exactly as they were presented, and the key that deanonymizes the randomized columns. For each line we take the frame at the temporal midpoint of its subtitle interval, so that the panel falls inside the delivery of the line rather than on a cut. The subtitle is composited bottom-centered in white with a dark outline, wrapped to at most two lines at a break a player would take, with the English source beneath it. Glyph size is scaled inversely with the number of columns, so that the printed subtitle has the same physical size regardless of how many systems a figure shows. Five columns of legibly rendered subtitle do not fit across the page width, so both figures are set landscape. No target text is edited, repunctuated, or rewrapped by hand.

Verification. SMART's translation of this episode was revised after the human study was run. Every SMART line in the figures was therefore checked character-for-character against the released translation of the episode, and only lines that are identical in both are used. One candidate contrast was discarded on this basis, because a later revision changed the very word it turned on. The five columns are thus text that is fixed and citable rather than a moving target compared against frozen baselines.

What the six panels show. Read together, the panels locate SMART's margin where the tables locate it. Four of them, (a), (b), (e), and (f), turn on reading the source correctly rather than on writing Chinese well, and two of those, (b) and (f), turn specifically on the material being tradecraft,

Table 22: The six lines of Figures 5 and 6, in panel order. Timecodes are episode time in Nikita S01E01; Passage indicates the ordinary-dialogue (A) or external-knowledge (B) passage of the human study; Dimension names the SubMQM dimension that the contrast bears on.

Panel	Timecode	Passage	English source line	Dimension
(a)	34:05	B	Makes one of us.	Accuracy
(b)	34:51	B	We trained Nikita to be a ghost.	Terminology
(c)	34:54	B	Finding her when she doesn't want to be found is next to impossible.	Fluency
(d)	25:14	A	The only reason why you're alive is because she wanted you that way.	Aud. Approp.
(e)	25:10	A	Yeah, was that before or after she duct-taped you to that springy rocking horse?	Accuracy
(f)	34:45	B	Black arrow was blown.	Terminology

which is the reading that content profiling supplies and that a single call has no reason to reach on a four-word line. The remaining two, (c) and (d), are lines that every system understands, where only SMART writes them the way a subtitle is written. The panels also show the limits of the illustration. They cover one direction, one genre, and one episode, and what they make visible are the semantic dimensions; they say nothing about the punctuation, capitalization, locale-formatting, and characters-per-line error types, which are localized and relatively sparse and for which SubMQM rather than inspection is the appropriate instrument.

I Reproducibility and Release

I.1 Released Artifacts

We release three artifacts. **(i) The Subtitle Arena construction pipeline**: the alignment, episode-keying, SRT reconstruction, and filtering code of Appendix D, together with the episode keys $(\mathrm{IMDb}, \mathrm{season}, \mathrm{episode})$ and the adaptation/inference split indices for every series, so that the benchmark can be rebuilt from the upstream corpus. **(ii) The SubMQM protocol**: the full 19-error-type rubric of Table 10, the judge prompts, the window and reference-alignment procedure, and the aggregation weights. **(iii) The SMART implementation**: agent prompts, the initial routing table, the tool interfaces of Table 8, and the evolved configurations (Π_T, ρ_T) for each series. All released artifacts are available at `https://anonymous.4open.science/r/8774/`.

I.2 Redistribution and Licensing

We do not redistribute subtitle text. Subtitle Arena is released as a reconstruction recipe over OpenSubtitles2024 (Tiedemann & Luo, 2026) rather than as a corpus: users obtain the upstream release themselves and our scripts regenerate the episode-level SRT streams and the splits from it. The OpenSubtitles2024 release is distributed under the Open Data Commons Attribution (ODC-BY) license. The upstream release redistributes subtitle text contributed to OpenSubtitles.org without claiming ownership of that content and maintains a takedown route for rights holders; our reconstruction artifacts retain attribution to the upstream source and do not include the subtitle text itself.

From IMDb we use only the title identifier, and only as a join key for grouping episodes of the same series under a canonical episode key. We do not redistribute IMDb metadata fields. The 14 genre categories reported in Subtitle Arena are released as benchmark-derived annotations: each series is assigned a single coarse category from the fixed taxonomy used in Table 1, rather than redistributing an upstream provider's genre metadata.

J Limitations and Future Work

Evaluation. SubMQM is an LLM-based protocol. We mitigate judge dependence by swapping the judge model (Appendix F.3) and by running a human study (Appendix F.7), but both checks cover a

subset of directions, and the Online subtitles we report alongside are community releases rather than a controlled human upper bound.

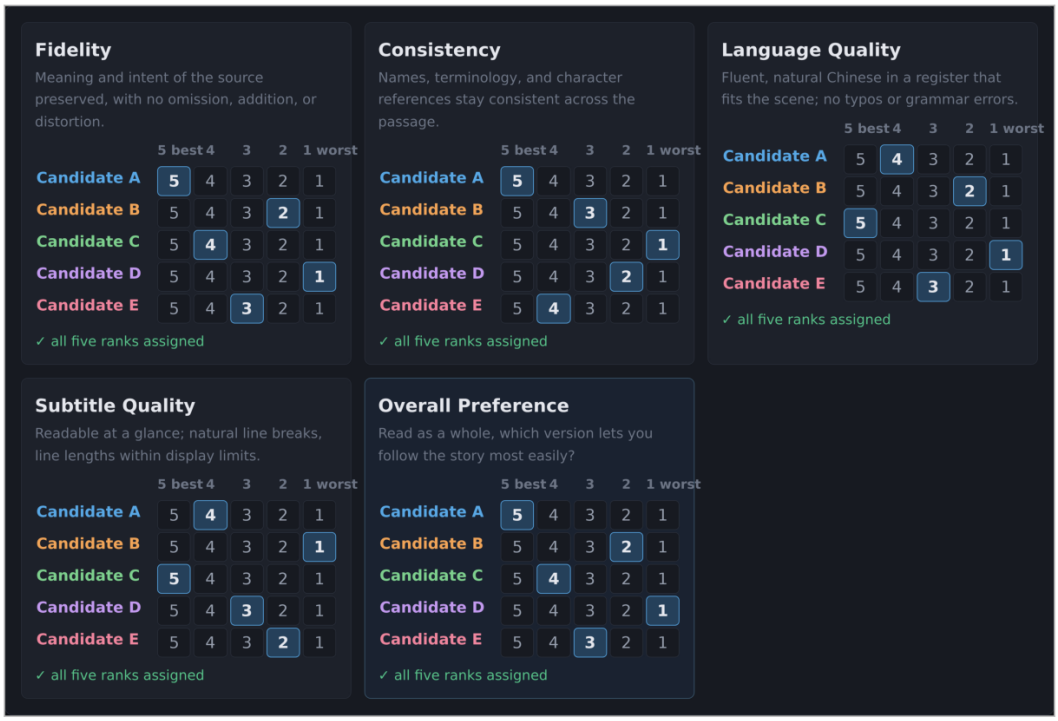
Cost. SMART spends substantially more test-time computation than a single-call translator (Appendix F.6), which limits its use in latency-bound or high-volume settings. Reducing the number of model calls without losing the consistency gains is the most direct extension.

Coverage. Subtitle Arena is derived from a single upstream corpus of pre-2024 subtitles and is English-pivoted, so non-English pairs and locales absent from that corpus remain untested; the temporal split of Appendix F.4 probes only the first of these. The qualitative analyses cover one direction and a small number of episodes.

Modality. Multimodal evidence is optional and is invoked per sentence (Appendix G); speaker diarization, longer-horizon visual context, and audio prosody are not yet part of the persistent series memory.



(a) Passage view: clip, English source, and the five anonymized candidates



(b) Ranking panel: four criteria plus overall preference, each ranked 5 to 1

Figure 3: The annotation interface used in the human study. **(a)** Each passage is presented with its video clip, the English source lines, and the five candidate translations as anonymized columns whose order is randomized per annotation instance. **(b)** Annotators rank the five candidates from 5 (best) to 1 (worst) on Fidelity, Consistency, Language Quality, and Subtitle Quality, and then on Overall Preference; a passage can be submitted only once every panel has all five ranks assigned. The passage shown is from Nikita S01E01.



(a) SMART (text only) (b) SMART + MM

Figure 4: A case in which visual evidence changes the translation, on the same frame for both systems. The magnifier shows a single mark on a bone, which the subtitle alone does not specify. Text-only SMART renders striation as a generic plural (**(a)**, 这些条纹), while SMART + MM queries the aligned video clip, recovers that one mark on a bone is in view, and renders it as a single bone striation (**(b)**, 这道骨纹). The English source is composited beneath the Chinese for reference.



(a) Idiom inversion — Nikita S01E01, 34:05. All four baselines read the sarcastic inversion of "that makes two of us" as inclusion; only SMART recovers that the speaker excludes himself.



(b) Metaphor preservation — Nikita S01E01, 34:51. Online deletes the ghost metaphor, and TransAgent and both single calls transfer it without anchoring it; SMART keeps it and marks it as tradecraft.



(c) Idiomatic compression — Nikita S01E01, 34:54. TransAgent and both single calls restate the conditional and overrun the reading-speed budget; SMART reaches a fixed expression of the same force within it.

Figure 5: Rendered subtitle comparison on Nikita S01E01, English→Chinese, panels (a) and (b). Each group states the phenomenon, the timecode, and the reading the comparison turns on, so that the panels can be read without this caption. Two observations the panels do not state. In (a), `Makes one of us` is a sarcastic inversion of "that makes two of us," and this is the one line in the set on which every system other than SMART agrees and is wrong. In (b), the TransAgent panel spells the protagonist's given name with a different character than TransAgent itself uses for that name in the other two passages of the study, whereas Online, Claude Opus 4.8, and SMART use one spelling throughout; cross-passage naming consistency is what series memory (Section 3.3) is for, and here it is an agentic pipeline rather than a single call that fails it.



(d) Register — Nikita S01E01, 25:14. Every baseline renders the causal clause literally; SMART replaces it with the conventional Chinese idiom for being spared.



(e) Source completeness — Nikita S01E01, 25:10. "Duct-taped" is dropped by Online and by the single call to Sonnet 4.6, SMART's own backbone, and kept by TransAgent, GPT-5.5 and SMART: on a fixed backbone the difference is the scaffold.



(f) Tradecraft terminology — Nikita S01E01, 34:45. "Blown" means compromised, not unsuccessful; Online reports a failure, TransAgent and both single calls use a general-purpose verb, and SMART the idiomatic Chinese one.

Figure 6: Rendered subtitle comparison on Nikita S01E01, English→Chinese, panels (c) and (d). Groups are annotated as in Figure 5. These are the two panels in the set on which no system misreads the source. What separates them is how the Chinese is written: (c) is a line every baseline renders as a restated conditional that exceeds the reading-speed budget for its duration, and (d) a line every baseline renders as a literal causal clause.