Style Diagnostic Profile for Translation: A Preliminary JA–EN Study

Yuri Amimoto, Ryosuke Mitani, Rem Hida, Ritvik Choudhary, Toshiyuki Sekiya
Sony Group Corporation, Tokyo, Japan
{yuri.amimoto, ryosuke.mitani, remu.hida,
ritvik.choudhary, toshiyuki.sekiya}@sony.com

Abstract

In Japanese–English machine-translation,especially for creative and localized content, standard MT metrics typically summarize translation quality as scalar scores, but provide little diagnostic information about how style changes in translation. This is limiting because politeness, directness, interpersonal stance, and character voice are encoded differently across Japanese and English, and stylistic change may reflect localization-oriented adaptation rather than error. We propose the Style Diagnostic Profile (SDP) framework: SDP assigns each utterance a seven-dimensional bipolar style vector, and Δ_{SDP}, the difference between target and source SDP vectors, represents translation-induced stylistic change as a multi-axis shift vector. To evaluate the framework, we first validate an LLM-based SDP scorer against bilingual human annotations on professional JA→EN subtitles and show that it approximates human style judgments. Using Style25, a controlled JA↔EN style-perturbation benchmark, we demonstrate that different perturbation categories occupy distinct regions in Δ_{SDP} space. As a case study, we analyze professional subtitles and observe stylistic adaptation rather than simple style copying. These results position the SDP framework as a diagnostic layer complementary to adequacy-oriented MT evaluation.

1 Introduction

For Japanese–English machine-translation, especially for creative and localized content such as subtitles, MT evaluation must consider not only adequacy and fluency but also how style changes across languages. Recent neural metrics such as COMETKiwi (Rei et al., 2022), MetricX-24 (Juraska et al., 2024), and GEMBA-DA (Kocmi and Federmann, 2023) are widely used to assess adequacy and fluency, but provide little diagnostic information about such stylistic changes. Style depends on interlocutor, situation, genre, and cultural convention (Brown and Levinson, 1987; Blum-Kulka, 1987): politeness, intimacy, authority, emotional stance, and directness are encoded differently in Japanese and English , and can shift independently of one another, so that a single scalar cannot capture which aspect changed. This is especially salient in Japanese–English creative translation, including anime, manga, games, and audiovisual media, where translators often adjust register, character voice, and interpersonal stance to satisfy localization policies, length constraints, genre expectations, and reader experience (Venuti, 1995; House, 2015; Díaz Cintas and Remael, 2007). Thus, stylistic change is not necessarily an error, but can be a natural and even necessary part of translation.

Source (JA).
いかん！お前は今死んではいかん！
(Noooo! You must not die now!)
SDP(s): pol= −2, emo= +2, dir= +1

Hyp A over-formalized; less preferred
"It would be inadvisable for you to die at this time."
Ours **SDP(t)**: pol= +1, emo= 0, dir= −1; **SFS**= 0.45
Baseline **COMETKiwi**= 0.85, **GEMBA-DA**= 0.82

Hyp B human-preferred
"Get away! You must not die now!"
Ours **SDP(t)**: pol= −2, emo= +2, dir= +1; **SFS**= 0.89
Baseline **COMETKiwi**= 0.82, **GEMBA-DA**= 0.85

Figure 1: **Motivating Japanese–English example.** Similar baseline scalar scores hide different stylistic shifts in a creative translation settings, which our SDP/SFS make explicit.

For this reason, the key question is not simply whether style is preserved, but how it shifts: whether a translated sentence became more polite, less emotional, more authoritative, or less direct. What is needed is a diagnostic representation of stylistic shift. We therefore propose the *Style Diagnostic Profile* (SDP) framework: SDP assigns each

Proceedings of the 13th Workshop on Asian Translation (WAT 2026), pages 17–45
November 9, 2026 ©2026 Association for Computational Linguistics

utterance a seven-dimensional bipolar style vector over interpretable axes, because stylistic shift is multi-faceted, a single scalar cannot indicate which dimension moved, and ΔSDP, the difference between target and source SDP vectors, identifies which stylistic dimensions changed and in which direction.

Figure 1 illustrates the motivation with a Japanese–English subtitle example. Two English translations of the same Japanese utterance receive similar scalar quality scores, yet human annotators prefer Hyp B. The difference is not primarily one of adequacy, Hyp A over-formalizes an urgent command, reducing emotional force and directness, whereas Hyp B better matches the source's interpersonal and affective profile. The SDP framework makes this contrast explicit, turning a vague sense that "the style is off" into an analyzable pattern.

Through scorer validation, controlled style perturbations, and a case study of professional JA→EN subtitles, we illustrate what an SDP layer can reveal about style in MT evaluation. We position SDP as a diagnostic layer alongside quality-oriented metrics, not as a new adequacy metric. Because it maps stylistic change onto interpretable axes, such a layer may also support human-in-the-loop review workflows (Wu et al., 2022), where making stylistic feedback explicit is valuable(§7).

Our contributions are as follows:

- We propose the SDP framework for Japanese–English style diagnostics: a per-utterance seven-axis bipolar style profile, and ΔSDP, a difference vector that identifies which dimensions changed and in which direction. We validate an LLM-based SDP scorer against bilingual human annotations on professional JA→EN subtitles (§3, §5).

- We introduce Style25, a controlled JA↔EN style-perturbation benchmark, showing that targeted perturbation categories occupy distinct regions in ΔSDP space, i.e., controlled evidence that SDP captures axis-specific stylistic shifts beyond scalar quality scores (§6.1).

- As a case study, we apply ΔSDP to professional JA→EN subtitles in Pro-JE and find evidence of systematic stylistic adaptation rather than simple style copying, with preliminary preference-relevant signal complementary to quality-oriented metrics (§6.2, §6.3).

2 Related Work

Quality-Oriented MT Metrics Widely used MT metrics, including reference-based measures such as BLEU (Papineni et al., 2002) and ChrF++ (Popović, 2017), learned metrics such as COMETKiwi (Rei et al., 2022) and MetricX-24 (Juraska et al., 2024), and LLM-based judges such as GEMBA-DA (Kocmi and Federmann, 2023), summarize translation quality as scalar scores. These metrics answer a different question from ours: whether a translation is good, not how its style changed and in which direction. We therefore treat these metrics as complementary quality-oriented filters rather than baselines for SDP to outperform: when adequacy or fluency is severely degraded, human judgments are dominated by quality problems rather than stylistic differences. §6.1 shows that, once adequacy is controlled, these metrics are largely insensitive to targeted stylistic shifts.

Style Preservation in Translation A separate line of work evaluates style directly. Style-transfer evaluation tests whether a target attribute, such as formality, changes while meaning is preserved (Rao and Tetreault, 2018; Briakou et al., 2021). In MT, RASTA (Havaldar et al., 2025) evaluates cross-cultural style alignment along politeness, intimacy, and formality, and reports that machine translation tends to neutralize stylistic variation, including translation to and from Japanese. Other work studies whether Big-Five personality traits survive Japanese–English translation (Fu et al., 2026), or evaluates style in translation in terms of semantic preservation, cultural specificity, and expression style (Tan et al., 2025). MQM-style frameworks can flag register or tone mismatches (Mariana et al., 2015; Li et al., 2025), but they mark that a stylistic problem exists rather than represent its signed direction. Our goal is therefore not to judge stylistic shifts as errors, but to represent their direction and magnitude at the utterance level.

Translation Studies Translation studies has long treated translation as a contextualized transformation for a new audience, rather than a literal transfer of meaning (House, 2015; Venuti, 1995; Reiss and Vermeer, 2013). This view is especially relevant to Japanese–English creative and subtitle translation, where character voice, register, and localization policy can motivate deliberate stylistic shifts (Díaz Cintas and Remael, 2007). SDP operationalizes this stance by describing how style moves along each

Axis	−2	0	+2
Politeness	Casual	Neutral	Formal
Roughness	Gentle	Neutral	Crude
Emotion	Flat	Neutral	Emotional
Authority	Submissive	Equal	Authoritative
Age	Childlike	Neutral	Mature
Confidence	Hedging	Neutral	Assertive
Directness	Indirect	Neutral	Direct

Table 1: Seven bipolar stylistic axes with pole labels. Functions: Politeness (social distance), Roughness (verbal aggression), Emotion (affective intensity), Authority (power direction), Age (age-graded language), Confidence (epistemic certainty), Directness (communicative directness). Each axis is scored on a five-point scale (-2 to $+2$) for both source and target utterances.

axis before judging whether the shift reflects adaptation, degradation, or preference-relevant variation.

3 Style Diagnostic Profile

3.1 Style Dimensions

To make stylistic movement explicit in Japanese–English translation, SDP represents each utterance as a seven-axis over $[-2, +2]$, with 0 as the neutral (Table 1). The signed value encodes both the direction and the magnitude of a stylistic property. The seven axes were developed through an inductive analysis of a 34,264-line Japanese–English anime subtitle corpus covering 37 works. After removing low-style-content utterances such as formulaic greetings, brief acknowledgments, and duplicates, we organized utterances that reveal character voice and interpersonal stance into recurrent bipolar contrasts. We then selected the final set of axes using four criteria: bipolar contrast, conceptual separability, cross-language applicability, and attestation across diverse types of anime and settings. We treat the axes as designed diagnostic contrasts for human-interpretable reporting. Appendix A.2 shows that empirical factors cut across the designed axes and that the axes exhibit no severe redundancy.

Because these axes are intended for diagnostic interpretation, they are defined as graded dimensions rather than categorical labels. For example, the Age axis does not predict chronological age; following Eckert (1997), it captures age-related linguistic style as a social position on a five-point scale from youthful to mature. Appendix A.5 provides worked examples of stylistically similar but distinct shifts.

3.2 Profile Representation

Given a source utterance s and its translation t, the scorer assigns each a seven-axis vector:

$$\mathrm{SDP}(x) = (\mathrm{SDP}_1(x), \ldots, \mathrm{SDP}_K(x)), \quad (1)$$

where $K = 7$ is the number of style axes and each component $x_k \in \{-2, -1, 0, +1, +2\}$ is the score on axis k. The diagnostic profile is the difference between target and source profiles:

$$\Delta_{\mathrm{SDP}}(s, t) = \mathrm{SDP}(t) - \mathrm{SDP}(s). \quad (2)$$

A zero component means that the source and translation receive the same score on that axis, whereas a non-zero component indicates a perceived style shift. The sign of the component specifies the direction of the shift.

3.3 Auxiliary Scalar Summary: Style Fidelity Score

SDP and Δ_{SDP} are the main representations in this work. However, some controlled perturbation experiments require a single scalar for ranking candidates. For this limited purpose, we define the Style Fidelity Score (SFS) as:

$$\mathrm{SFS}(s, t) = 1 - \frac{1}{K} \sum_{k=1}^{K} \frac{|\Delta_{\mathrm{SDP}_k}(s, t)|}{R}, \quad (3)$$

where $R = 4$ is the maximum possible absolute difference between two scores on the $[-2, +2]$ scale. Thus, SFS $\in [0, 1]$. SFS is a lossy scalar summary of stylistic closeness, not a measure of translation quality or translation appropriateness. Professional subtitle translation may deliberately adapt style to the target language or audience (§6.2), and such adaptation can lower SFS even when the translation is appropriate.

4 Experimental Setup

This section describes the datasets, human annotations, metrics, baselines, and evaluation protocols used in our experiments. We use three Japanese–English subtitle resources for complementary evaluations: Subtitle-100 for scorer validation, Style25 for controlled JA↔EN style perturbations, and Pro-JE for analyzing how SDP dimensions and SFS relate to human preference. All human annotations were conducted by paid in-house bilingual

Japanese–English annotators, not crowdworkers. To clarify the data contribution: the underlying utterances come from existing resources (professional anime subtitles for Subtitle-100 and Pro-JE; OpenSubtitles (Lison and Tiedemann, 2016) for Style25, not redistributed), whereas the style-perturbed variants (in Subtitle-100 and Style25), the Pro-JE LLM candidates, and all seven-axis SDP and BWS annotations are newly created through our in-house bilingual workflow.

We compare three metrics families. The SDP family consists of SDP, a seven-axis diagnostic profile, and SFS, a scalar style-fidelity score computed from ΔSDP. SA-3 is a controlled three-axis ablation based on the three style axes used in RASTA (Havaldar et al., 2025), using the same scorer and distance-to-fidelity transformation as SFS.[1] Quality-oriented baselines include COMETKiwi, MetricX-24, and GEMBA-DA[2].

4.1 Subtitle-100: Scorer Validation

Subtitle-100 contains 100 professional JA→EN anime subtitle pairs. Three bilingual annotators, masked to the condition, independently rate each pair on all seven axes using the five-point bipolar scale (36.5 annotator-hours total). Of the 100 pairs, 25 retain the original professional subtitle (*Official* condition) while the remaining 75 use stylistically modified versions (*Modified* condition), created by rewriting the official subtitle along one or more axes (Appendix A.1); the larger Modified set provides more perturbed items for the SFS separation test, while the Official set anchors scorer validation against unmodified professional style (Appendix D).

4.2 Style25: Controlled Style Perturbations

Style25 contains 200 source utterances: 100 items in each direction for JA⇔EN subtitle items sampled from OpenSubtitles, used to obtain broad subtitle dialogue beyond anime for automatic sensitivity tests; the original subtitle text is not redistributed. For each seed translation, we use Gemma3-27B-IT (Team et al., 2025) to rewrite it into 25 target styles, grouped into 9 categories (Table 9), at two intensity levels via a one-shot style-conditioned prompt (Appendix B.2). To keep candidates comparable in adequacy, each rewritten variant must pass a COMETKiwi quality gate (≥ 0.80); variants below the threshold are regenerated when possible or otherwise discarded, so that surviving candidates differ from the unmodified translation (the anchor) mainly in style rather than meaning.

For each item, we rank the anchor together with its style-perturbed variants, and define Anchor-Recovery Preservation@1 (Pres@1) as the fraction of items for which a metric ranks the anchor first. Since the anchor is a context-aware translation tuned to the work's overall tone, it is expected to be closer to the source style than any single-style variant, so a style-sensitive metric should favor it on average. We therefore read Pres@1 as a metric's aggregate sensitivity to controlled style shifts, not as a measure of quality, preference, or ideal style copying; it need not reach 1.0, since a perturbed variant may occasionally align with the source style by chance.

The primary comparison is SFS vs. SA-3, which share the same scorer and distance-to-fidelity transformation and thus isolate the effect of the axis set. MetricX-24 and GEMBA-DA serve as quality-oriented baselines; COMETKiwi is excluded as a selector, as it is already the generation-time quality gate.

4.3 Pro-JE: Human Preference in Professional Subtitles

Pro-JE contains 32 professional JA→EN anime subtitle items, separate from Subtitle-100. Each item has four candidates: the official human subtitle and outputs from three LLMs (Qwen3-14B-Instruct (Yang et al., 2025), Gemma3-12B-IT, and Llama3.1-8B-Instruct (Grattafiori et al., 2024); Appendix B.3). Four bilingual annotators judged them using Best–Worst Scaling (BWS; Louviere et al., 2015). We use Pro-JE as a case study of how particular SDP dimensions and SFS relate to human preference, not to argue that style preservation is generally preferred.

We compare metric-induced rankings against human BWS preferences, reporting Kendall's τ and Top-1 accuracy (with fractional credit for ties; Appendix F). For scalar metrics, higher scores indicate either better predicted quality (COMETKiwi, MetricX-24, GEMBA-DA) or greater source-style closeness used diagnostically (SFS, SA-3); SDP itself is a seven-axis diagnostic profile, not a scalar ranking score.

[1] SA-3 is not a reproduction of RASTA, but an axis-set ablation under our scoring framework.

[2] We use a reference-free DA-style prompt that asks an LLM for a single 0–100 quality score given the source and candidate

5 Validating the LLM-based SDP Scorer

Because all downstream SDP analyses depend on reliable axis assignment, and LLM-as-a-judge reliability is not guaranteed (Zheng et al., 2023), we validate candidate scorers against human annotations before applying SDP in our diagnostics.

We use Subtitle-100, which contains human annotations for 100 professional JA→EN anime subtitle pairs on all seven SDP axes. Each source and target utterance was scored separately on each axis ($100 \times 2 \times 7 = 1{,}400$ labels), and the scorer was queried once per utterance–axis pair using the same five-point axis rubric. For the human reference, three bilingual annotators independently scored each item; we averaged their scores and rounded to the nearest integer on the five-point scale to obtain the final label.

Human agreement is moderate overall, but varies by axis: authority is relatively reliable, whereas age and directness are weaker. The mean Krippendorff's α across the seven axes is 0.611 for Japanese and 0.492 for English (overall mean 0.551), comparable to prior work on pragmatic and social meaning annotation (Pei and Jurgens, 2020). Detailed inter-annotator agreement and axis-level validation results are reported in Appendix C.

Table 2 compares three open-weight candidate scorers from different model families and scales, together with two non-adaptive baselines against human labels. We report five agreement metrics: exact-match accuracy, within-one (±1) accuracy on the five-point scale, mean absolute error (MAE), root mean squared error (RMSE), and the mean per-axis Spearman correlation (ρ). Qwen3-14B-Instruct (Yang et al., 2025) achieves the best overall agreement with human annotations: exact 55.6%, within-one 93.9%, MAE 0.51, RMSE 0.81, and mean per-axis Spearman ρ=0.47. The all-zero baseline always predicts the neutral label, while the majority-vote baseline predicts the most frequent human label for each axis. Because both baselines make constant predictions, their Spearman correlations are undefined.

We therefore use Qwen3-14B-Instruct as the SDP scorer in the downstream analyses. This validation supports using its outputs as automatic style diagnostics, while results involving weaker axes such as age and directness should be interpreted cautiously. The scorer also shows a systematic zero bias relative to humans: it outputs the neutral label 0 more often than human annotators, with the largest gap on age and politeness (Appendix C.7). This bias should be kept in mind when interpreting near-zero ΔSDP values as an absence of style shift.

Model	Exact↑	±1↑	MAE↓	RMSE↓	ρ ↑
Qwen3-14B-Instruct	**55.6**	**93.9**	**0.51**	**0.81**	**0.47**
GPT-oss-120B	46.1	88.4	0.69	1.03	0.37
Gemma3-9B	48.9	87.5	0.67	1.04	0.39
All-zero baseline	45.5	88.6	0.66	0.94	n/a
Majority-vote baseline	49.8	88.9	0.63	0.95	n/a

Table 2: LLM SDP scorer validation on Subtitle-100. Exact and ±1 accuracy are percentages; ρ is the mean per-axis Spearman correlation. Constant baselines have undefined ρ due to zero prediction variance.

6 Results and Analysis

Using the validated SDP scorer (§5), we report three complementary analyses. §6.1 uses the controlled Style25 benchmark to test *whether ΔSDP is diagnostically sensitive to deliberate style perturbations*, i.e., whether perturbation categories occupy distinct regions in ΔSDP space. §6.2 applies ΔSDP to professional Pro-JE subtitles to examine *how style is adapted in practice*. §6.3 relates these shifts to human preference to ask *which dimensions carry preference-relevant signal*. We treat Style25 as primary controlled evidence and Pro-JE as a case study.

6.1 Controlled JA↔EN Style Sensitivity

We use Style25 to test whether perturbation categories produce category-specific shifts in ΔSDP space, and whether these shifts differ between JA→EN and EN→JA. The full seven-axis ΔSDP is the diagnostic space; for visualization, axes that move together under a common stylistic direction can be aggregated into one interpretable dimension. Figure 2 uses one such two-dimensional view:

$$\textit{interpersonal force} = \text{mean}(\Delta\text{dir}, \Delta\text{aut}, -\Delta\text{pol}),$$

$$\textit{expressive intensity} = \text{mean}(\Delta\text{emo}, \Delta\text{rou}, \Delta\text{con}).$$

These two axes are an exploratory visualization motivated by the interpersonal and affective characteristics that are especially salient in anime dialogue. Age is omitted here as it contributes little discriminative signal for these categories, but remains part of the full representation. This projection is only a task-specific visualization: the category separation is not an artifact of it, as a data-driven PCA on the full seven-axis ΔSDP reproduces the same structure,

with the first two components explaining most of the variance (Appendix H).

The 25 individual styles are grouped into nine categories. The community-register category has the largest expressive-intensity shift, while pragmatic strategy stays near the origin. The anchor is not at the origin, indicating that naturally occurring translations can already involve cross-lingual stylistic shifts rather than being style-perfect.

Anchor recovery as an auxiliary diagnostic. As an auxiliary diagnostic, we test whether each metric ranks the anchor above its style-perturbed variants (Pres@1, Rank@3, Rank@5; §4.2); high recovery indicates sensitivity to controlled perturbations, not higher quality or human preference.

Table 3 shows that anchor recovery depends on metric type and direction. Both style metrics exceed chance (5.7% for JA→EN, 3.7% for EN→JA) by a wide margin. SFS outperforms SA-3 in JA→EN, whereas SA-3 is slightly higher in EN→JA, suggesting that the most useful style cues differ by direction. GEMBA-DA appears strong under default ranking, but largely due to ties from heavily discretized scores (Appendix D), while MetricX-24 rarely selects the anchor. These results characterize metric behavior under controlled perturbations, not human preference for the anchor.

Category-specific sensitivity Anchor recovery also varies by perturbation category (Table 18 in Appendix E.4). SFS is stronger than SA-3 for categories involving dimensions beyond formality and politeness, such as dialect, locale, character voice, and speech style. SA-3 is competitive for categories more directly tied to formality or politeness, such as register and stance. This pattern is consistent with the seven-axis design: broader axes help when perturbations affect multiple aspects of character voice, while a smaller formality-oriented set is sufficient for narrower shifts.

6.2 Professional Style Adaptation in Pro-JE

This case study asks how professional translators adapt style in practice. For the 32 professional JA→EN anime subtitles in Pro-JE, we score source and target on all seven axes with the validated scorer, compute the per-axis mean ΔSDP (target minus source), and read each axis as preserved (near-zero shift) or adjusted (Table 4).

Emotion and directness remain nearly unchanged (average shifts -0.02 and $+0.03$), suggesting that translations preserve the affective and structural core of the utterance. In contrast, interpersonal register is adjusted: politeness ($+0.24$), authority ($+0.19$), and roughness ($+0.17$) all move toward a more neutral target profile.

These changes are consistent with the nature of anime dialogue, where reduced-politeness or very gentle soft source-side expressions are rendered with a more neutral tone. Overall professional subtitle translation in this sample preserves emotion and directness while normalizing some interpersonal register cues for the target language. A broader comparison with controlled Style25 perturbation profiles is provided in Appendix E.3.

6.3 Relation to Human Preference

As a secondary check on Pro-JE, we relate style shift to human preference on a set that contains professional JA→EN subtitles and outputs from three LLMs, together with human BWS judgments. We use this setting not to claim that source-style preservation is always preferred, but to examine which style dimensions carry preference-relevant signal.

On average, the Human translation is most preferred (group-mean BWS $+0.358$; Table 26 in Appendix I). It also shows the smallest emotion shift ($|\Delta|$=0.22), while Llama, which has the largest directness shift ($|\Delta|$=1.07), ranks last (-0.345). At the same time, the contrast is not absolute: the LLM outputs are often fluent and adequate, and some items do not yield a clear human-translation advantage. This makes Pro-JE a useful but limited setting for testing whether style contributes information beyond standard quality signals.

A representative Pro-JE item illustrates how ΔSDP can explain preference patterns missed by scalar quality scores. For the source utterance すげえ 野望のサイズがデカい (“the scale of that ambition is huge”), COMETKiwi ranks the Qwen3-14B-Instruct candidate above the official subtitle (0.83 vs. 0.79), whereas human BWS strongly prefers the official subtitle ($+1.00$ vs. 0.00). The ΔSDP profile shows that both translations reduce roughness and directness, but the Qwen3-14B-Instruct candidate additionally raises politeness and confidence ($+1$ each), making the line more standard and assertive than the source character voice. The full example is shown in Table 25.

SFS shows a modest positive association with human preference using nominal significance thresholds ($\tau = +0.147$, uncorrected $p < .05$). This association appears complementary to quality met-

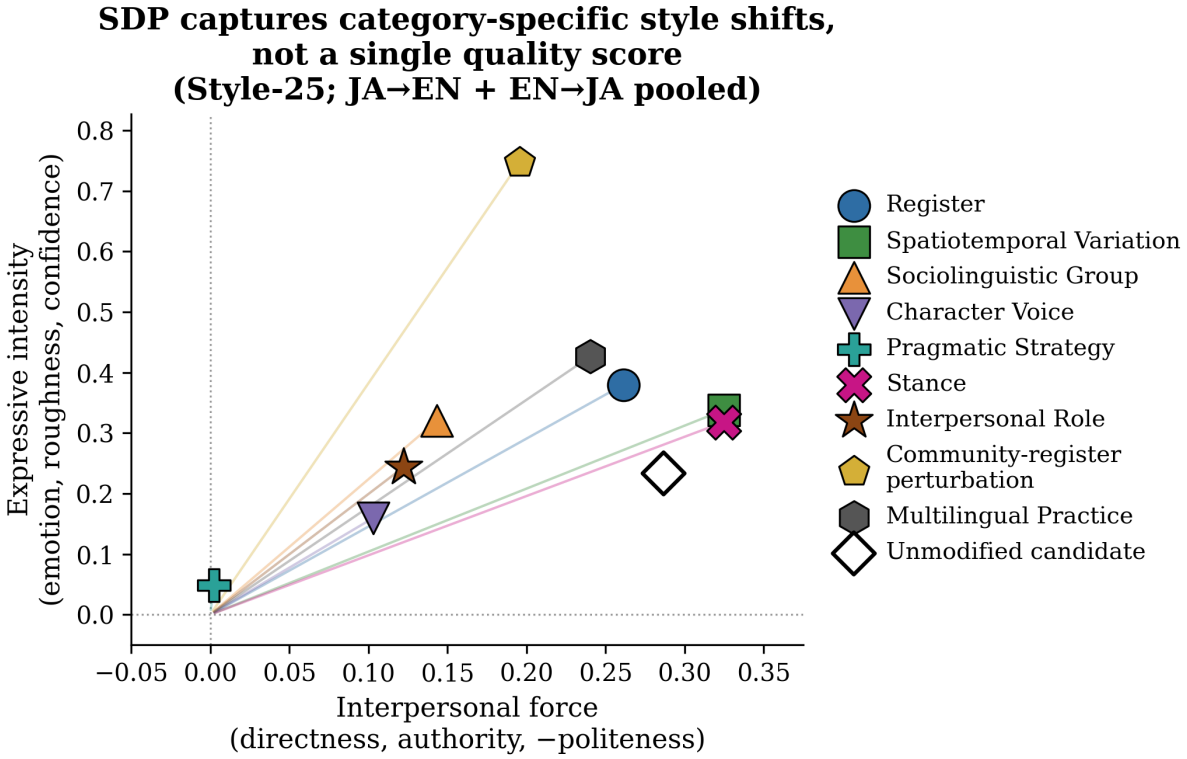


Figure 2: **Interpersonal–expressive style plane (Style25, JA↔EN pooled).** Each point is a Style25 category placed by its mean ΔSDP after projecting the seven-axis representation onto two visualization axes; the white diamond is the unmodified OpenSubtitles anchor. Categories occupy distinct regions, indicating that ΔSDP captures category-specific stylistic shifts rather than collapsing them into a single quality score.

rics: on Pro-JE, COMETKiwi itself is uncorrelated with BWS ($\tau \approx -0.008$), while the SFS association strengthens after quality gating ($\tau = +0.225$, uncorrected $p < .05$). GEMBA-DA itself attains a comparable or slightly higher τ on this small sample ($\tau = +0.160$, uncorrected $p < .05$; Appendix F.1); none of these associations survive Holm–Bonferroni correction for multiple comparisons (Appendix F.1, F.2), so we treat them as case-study observations rather than confirmatory findings. SFS's value lies in explaining preference through an interpretable, quality-independent axis rather than in exceeding quality-metric correlation. We interpret this as case-study evidence that source-style closeness can carry preference-relevant signal in some cases, not that style preservation is generally optimal. Full per-metric correlations and quality-gate analyses are reported in Appendix F.

7 Discussion

SDP should be interpreted as a diagnostic complement to quality-oriented MT metrics, not a replacement: metrics such as COMETKiwi, MetricX-24, and GEMBA-DA provide scalar quality judgments, whereas ΔSDP describes how style moves across axes.

We are deliberately careful to separate what our evidence supports at different strengths. Style25 is a controlled, automatic test at scale, and it shows that perturbation categories produce category-specific shifts in ΔSDP space rather than collapsing into a single quality-like dimension. The Pro-JE preference result is intentionally narrower: on 32 items, style fidelity carries preference-relevant signal complementary to scalar quality scores, but the associations are uncorrected and do not survive multiple-comparison correction (§6.3). We therefore treat Style25 as our primary controlled evidence for the SDP framework's diagnostic sensitivity and Pro-JE as a case study, not as evidence that source-style preservation is generally optimal.

A further potential use case is human-in-the-loop translation review. Entertainment localization often involves multi-person workflows, where translators, editors, and quality-control reviewers may differ in their access to source-language nuance. SDP could help make otherwise vague stylistic feedback more explicit by mapping perceived changes onto

Metric	JA→EN			EN→JA			Type
	Pres@1	Rank@3	Rank@5	Pres@1	Rank@3	Rank@5	
GEMBA-DA	74.0	77.0	81.0	36.0	40.0	41.0	quality
SFS (7-axis)	40.0	57.0	69.0	20.0	35.0	42.0	style
SA-3 (RASTA)	31.0	52.0	62.0	24.0	40.0	47.0	style
MetricX-24	4.0	19.0	31.0	0.0	3.0	4.0	quality

Table 3: Anchor-recovery behaviour (Pres@1, Rank@3, Rank@5, %) on Style25. Pres@1/Rank@3/Rank@5 measure how often each metric ranks the unmodified OpenSubtitles anchor first, within the top 3, or within the top 5, respectively, among its deliberately style-perturbed variants. These are diagnostics of sensitivity to controlled perturbations, not a quality ranking, a human-preference result, or a claim that style copying is desirable. Chance for Pres@1 is approximately 5.7% for JA→EN and 3.7% for EN→JA, based on the average number of candidates after filtering (approximately $3\times$ and $5\times$ these values for Rank@3 and Rank@5, respectively). COMETKiwi is used as a generation-time quality gate and is therefore not included as a selector.

Axis	Src (JA)	Tgt (EN)	Shift
Politeness	-0.21	$+0.03$	$+0.24$
Authority	-0.11	$+0.08$	$+0.19$
Roughness	-0.45	-0.28	$+0.17$
Age	$+0.36$	$+0.24$	-0.12
Confidence	$+0.39$	$+0.47$	$+0.08$
Directness	$+0.32$	$+0.35$	$+0.03$
Emotion	$+0.33$	$+0.31$	-0.02

Table 4: Professional subtitle adaptation by axis in Pro-JE (Human, $N{=}32$). Emotion and directness show near-zero average shifts, while politeness, authority, and roughness show larger interpersonal-register adjustments.

interpretable axes, such as politeness, directness, and emotional intensity. This suggests a possible role for SDP as an orientation aid for post-editing and review, although validating such use requires dedicated user studies.

8 Conclusion

This paper asked whether stylistic change in translation can be made visible and analyzable, rather than collapsed into a single preservation or quality score. We introduced the SDP framework, a seven-axis per-utterance style profile together with ΔSDP, its signed source–translation style-shift vector, and Style25, a controlled JA↔EN style-perturbation benchmark. Together, these tools allow stylistic shifts to be diagnosed in terms of which dimensions changed and in which direction.

Our analysis suggests that ΔSDP captures category-specific stylistic shifts that scalar quality metrics cannot express. In our Pro-JE sample, professional JA→EN subtitles show evidence of systematic style adaptation rather than simple style copying. We also find uncorrected, case-study-level evidence that SFS carries preference-relevant signal after quality filtering, suggesting that style-sensitive signals can complement adequacy-oriented metrics. Throughout, our aim is descriptive rather than prescriptive: SDP is designed to show *how* style moved along each axis, not to judge whether that movement is inherently good or bad.

These findings suggest that signed style profiles can support more interpretable evaluation of Japanese–English creative and localized translation, where stylistic shifts may be acceptable or even desirable. Future work should improve LLM-based SDP scorer calibration, extend validation to other Asian-language pairs, and develop context-aware SDP scoring. The present scorer is intentionally lightweight, relying only on source–target utterance pairs, but stylistic acceptability in subtitles may depend on broader discourse context, speaker relationships, visual cues, and source-side vocal performance. We also plan to explore ΔSDP as structured feedback for post-editing or localization review.

Limitations

This study focuses on Japanese–English creative translation, especially subtitle-style data. The seven SDP axes were developed from Japanese–English anime subtitles, so other domains, genres, or language pairs may require different axes or subsets.

SDP is context-limited: scores are assigned from source–target utterance pairs without surrounding discourse, speaker identity, addressee information, scene context, or subtitle length constraints. Interpersonal axes such as politeness, authority, and directness therefore reflect perceived utterance-level stance, not verified claims about speaker relationships or the full communicative situation.

The LLM-based SDP scorer is not a substitute

for human judgment. It can attenuate shift magnitude, especially on axes with lower human–LLM agreement such as age and directness. Relatedly, our scorer selection is itself limited in scope: we compare only openly available models, motivated by reproducibility and computational-resource constraints, and leave an exhaustive search over scorers, including proprietary or larger models, to future work.

Finally, the human preference analysis is case study and Style25/Pres@1 should be interpreted as diagnostic sensitivity tests rather than translation-quality measures.

Ethics Statement

All human annotations were produced by professional bilingual Japanese–English data annotators through our in-house workflow rather than crowdworkers; they were informed of the research use and compensated as part of their regular paid work (§4). Style25's synthetic community-register condition (§4.2) is a controlled stress test, not a depiction of any real community, identity group, or protected category.

References

Shoshana Blum-Kulka. 1987. Indirectness and politeness in requests: Same or different? *Journal of Pragmatics*, 11(2):131–146.

Eleftheria Briakou, Sweta Agrawal, Joel Tetreault, and Marine Carpuat. 2021. Evaluating the evaluation metrics for style transfer: A case study in multilingual formality transfer. In *Proceedings of EMNLP*, pages 1321–1336.

Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage*. Cambridge University Press.

Jorge Díaz Cintas and Aline Remael. 2007. *Audiovisual Translation: Subtitling*. St Jerome.

Penelope Eckert. 1997. Age as a sociolinguistic variable. In Florian Coulmas, editor, *The Handbook of Sociolinguistics*, pages 151–167. Blackwell.

Markus Freitag, Nitika Mathur, Daniel Deutsch, Chi-Kiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchicchio, Chrysoula Zerva, and Alon Lavie. 2024. Are LLMs breaking MT metrics? results of the WMT24 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81, Miami, Florida, USA. Association for Computational Linguistics.

Xinger Fu, Masaaki Nagata, and Chenhui Chu. 2026. A study on speaker individuality preservation in English–Japanese translation. In *Proceedings of the 32nd Annual Meeting of the Association for Natural Language Processing*, pages 4162–4167. In Japanese.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Shreya Havaldar, Adam Stein, Eric Wong, and Lyle Ungar. 2025. Towards style alignment in cross-cultural translation. In *Proceedings of ACL*, pages 32213–32230.

Juliane House. 2015. *Translation Quality Assessment: Past and Present*. Routledge.

Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. MetricX-24: The Google submission to the WMT 2024 metrics shared task. In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504, Miami, Florida, USA. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203.

Yunmeng Li, Jun Suzuki, Makoto Morishita, Kaori Abe, and Kentaro Inui. 2025. MQM-chat: Multidimensional quality metrics for chat translation. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 3283–3299, Abu Dhabi, UAE. Association for Computational Linguistics.

Pierre Lison and Jörg Tiedemann. 2016. OpenSubtitles2016: Extracting large parallel corpora from movie and TV subtitles. In *Proceedings of LREC*, pages 923–929.

Jordan J. Louviere, Terry N. Flynn, and A. A. J. Marley. 2015. *Best–Worst Scaling*. Cambridge University Press.

Valerie Mariana, Troy Cox, and Alan Melby. 2015. The multidimensional quality metrics (MQM) framework: A new framework for translation quality assessment. *The Journal of Specialised Translation*, 23:137–161.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Jiaxin Pei and David Jurgens. 2020. Quantifying intimacy in language. *Preprint*, arXiv:2011.03020.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may I introduce the GYAFC dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of NAACL-HLT*, pages 129–140.

Ricardo Rei, Marcos Treviso, Nuno M. Guerreiro, Chrysoula Zerva, Ana C Farinha, Christine Maroti, José G. C. de Souza, Taisiya Glushkova, Duarte Alves, Luisa Coheur, Alon Lavie, and André F. T. Martins. 2022. CometKiwi: IST-unbabel 2022 submission for the quality estimation shared task. In *Proceedings of the Seventh Conference on Machine Translation (WMT)*, pages 634–645, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Katharina Reiss and Hans J. Vermeer. 2013. *Towards a General Theory of Translational Action: Skopos Theory Explained*. Routledge.

Xin Tan, Bowei Zou, and AiTi Aw. 2025. A benchmark for translations across styles and language variants. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 2389–2402, Suzhou, China. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Lawrence Venuti. 1995. *The Translator's Invisibility: A History of Translation*. Routledge.

Xingjiao Wu, Luwei Xiao, Yixuan Sun, Junhang Zhang, Tianlong Ma, and Liang He. 2022. A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 135:364–381.

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 46595–46623.

The appendices are organized as follows. Appendix A provides full axis definitions, the degeneracy analysis motivating the seven-axis design, and worked diagnostic case examples; Appendix B and Appendix C detail the datasets and LLM scorer validation; Appendix D and Appendix E.1 report additional statistical analyses on Style25; Appendix F reports the full Pro-JE preference and quality-gate analysis; Appendix G presents direction-specific style planes; and Appendix I presents complete case studies.

A Axis Definition

A.1 Full Axis Definitions with Anchor Examples

Each axis is scored on a five-point bipolar scale (-2 to $+2$). Table 5 provides the scale-point descriptions used during annotation and LLM scoring. Table 6 shows bilingual anchor utterances for one representative axis (politeness); anchors for all axes follow the same structure and are included in the LLM prompt (Appendix C.5).

A.2 Degeneracy Analysis: Why 7 Axes

To assess whether the 7 axes are redundant, inter-axis correlations and Variance Inflation Factors (VIF) were examined from human annotation data ($N = 100$ items, 3-annotator means). VIF values are reported in Table 7; all are below 4.0, indicating acceptable collinearity.

Inter-axis correlation structure. Three interpretable clusters emerge: (1) **Dominance cluster**: roughness–authority ($\rho = .75/.70$) and authority–confidence ($\rho = .54/.50$). (2) **Politeness–roughness**: inversely correlated ($\rho = -.56/-.48$) (Brown and Levinson, 1987). (3) **Peripheral axes**: age and directness show low to moderate correlation. Correlated axes are not merged: a character can be *authoritative yet gentle*, which a merged axis could not represent.

A.3 Empirical Validation of Three-Layer Structure

To validate the conceptual three-layer organization (Table 8), we conducted exploratory factor analysis, hierarchical clustering, and reliability analysis on human-annotated Subtitle-100 data (N=100 utterances, 3 annotators).

Correlation matrix. Spearman correlations reveal heterogeneous within-layer relation-

Axis	Function	-2	-1	0	$+1$	$+2$
Politeness	Social distance	Very Casual: zero social distance; friends/family	Casual: close but with consideration; relaxed settings	Neutral: standard social distance; public	Formal: clear hierarchy; business	Very Formal: maximum respect; ceremonies
Roughness	Verbal aggression	Very Gentle: comforting; tender word choice	Gentle: calm tone; respectful	Neutral: objective; neither aggressive nor soft	Somewhat Rough: warning/irritation; mild reprimands	Very Rough: gangster/pirate speech; vulgar language
Emotion	Affective intensity	Very Restrained: robot-like; textbook reading	Somewhat Restrained: calm professional; news reporting	Neutral: standard; moderate expression	Somewhat Emotional: recognizable emotion (joy, anger)	Very Emotional: screaming; extreme excitement
Authority	Power direction	Very Submissive: absolute obedience; pleading	Somewhat Submissive: subordinate; polite requests	Equal: colleagues; mutual exchange	Somewhat Authoritative: teacher; giving instructions	Very Authoritative: warden/dictator; commands
Age	Age-graded language	Very Childlike: baby talk; language errors	Youthful/Teen: teen slang; SNS-style	Neutral: standard adult; no age markers	Mature: professional adult speech	Very Mature/Archaic: historical drama; elderly speech
Confidence	Epistemic certainty	Very Uncertain: timid; avoids assertions	Somewhat Uncertain: cautious; some hedging	Neutral: balanced; fact-based	Somewhat Confident: expert opinion; clear	Very Confident: absolute claims; no doubt
Directness	Communicative directness	Very Indirect: heavy hedging; implies rather than states; avoids the point	Somewhat Indirect: softened suggestions; qualified requests	Neutral: states the point plainly, without extra softening or bluntness	Somewhat Direct: clear statements; direct requests	Very Direct: blunt imperatives; explicit commands; no hedging

Table 5: Seven bipolar stylistic axes with full scale-point descriptions.

ships (pooled across JA source and EN target):**Interpersonal**: Roughness–Authority strongly correlate ($r = 0.73$), but Politeness negatively correlates with both Roughness ($r = -0.50$) and Authority ($r = -0.48$).**Affective-Epistemic**: Emotion–Confidence show weak positive correlation ($r = 0.22$). **Expressive-Structural**: Age–Directness show weak *negative* correlation ($r = -0.35$).

Exploratory factor analysis (3 factors, oblimin rotation). The empirical factor structure differs from the three-layer organization, consistent with the orthogonal axis design: **Factor 1** ("Interpersonal Forcefulness"): high loadings on Roughness ($+0.92$), Authority ($+0.94$), and Politeness (-0.73). **Factor 2** ("Maturity & Assertiveness"): high loadings on Age ($+0.77$) and Confidence ($+0.72$), crossing layer boundaries. **Factor 3** ("Linguistic Simplicity"): dominated by Directness ($+0.96$) with secondary Emotion loading ($+0.44$). Variance explained: 75.4% (Japanese), 74.9% (English).

Reliability analysis (Cronbach's α). Cronbach's α reflects the designed axis orthogonality: **Interpersonal**: $\alpha = -0.39$ (Japanese), -0.35 (English); negative values result from inverse correlations (Politeness vs. Roughness/Authority). **Affective-Epistemic**: $\alpha = 0.34$ (Japanese), 0.22 (English); low but positive, reflecting weak correlation between independent dimensions. **Expressive-Structural**: $\alpha = -0.92$ (Japanese), -1.34 (English); Age and Directness measure independent aspects.

Interpretation. The low and negative α values validate the intended orthogonal design. The three-layer organization serves as a conceptual scaffold for interpretation, not a claim of empirical factor cohesion.

Score	Japanese	English
-2	やー、パス	Nah, can't
-1	ごめん、ちょい難しいかも	Sorry, I can't make it.
0	ごめんなさいね、難しそうです	I have a slightly different view.
$+1$	大変申し訳ございませんが、お引き受け致しかねます	I sincerely apologize, but I'm unable to accommodate that.
$+2$	誠に恐縮ではございますが諸般の事情によりお受けいたしかねます	I deeply regret that circumstances prevent me from accepting.

Table 6: Bilingual anchor examples for politeness (representative). Anchors for all seven axes follow the same structure.

Axis	VIF (JA)	VIF (EN)
politeness	3.01	3.62
roughness	2.78	3.25
emotion	1.70	1.38
authority	3.69	3.26
age	1.50	2.32
confidence	1.51	1.40
directness	1.30	1.87

Table 7: Variance Inflation Factors (full $N = 100$ human-annotated items). All below 4.0 (conventional threshold: 5.0), indicating low collinearity among the seven axes.

Layer	Axes	Function
Interpersonal	Pol, Aut, Rou	Social relationship constraints
Affective Epistemic	Emo, Con	Speaker emotional and epistemic state
Expressive Structural	Age, Dir	Sociolinguistically marked variation

Table 8: Three-layer organization of stylistic axes. Pol = Politeness, Aut = Authority, Rou = Roughness, Emo = Emotion, Con = Confidence, Dir = Directness.

A.4 Example of Profile Representation

For illustration, consider the Japanese source おおいやだ毛虫じゃあるまいし and its professional translation "I ain't a caterpillar!". Suppose the scorer assigns the following non-zero SDP(s) and SDP(t) components:

$$s: \{\text{pol:}-1,\ \text{rou:}+1,\ \text{emo:}+1,\ \text{aut:}-1,\ \text{con:}+1,\ \text{dir:}+1\},$$
$$t: \{\text{pol:}-2,\ \text{rou:}+1,\ \text{age:}-1\}.$$

Axes omitted from the display are treated as zero. The resulting diagnostic profile is:

$$\Delta_{\text{SDP}}(s,t) = \{\text{pol:}-1,\ \text{emo:}-1,\ \text{aut:}+1,\ \text{age:}-1,\ \text{con:}-1,\ \text{dir:}-1\}.$$

Here, zero components are omitted. Non-zero components indicate axes on which the translation differs stylistically from the source. Thus, ΔSDP identifies where stylistic nuance changes, while the sign of each component indicates the direction of that change. A scalar score cannot express this axis-level information.

A.5 Diagnostic Case Examples

This section provides extended diagnostic case examples illustrating how per-axis human annotation captures specific stylistic deviations. Each case reports three rows on the seven SDP axes: the **Source** (Japanese source utterance, scored directly), the **Official** translation, and a stylistically **Modified** variant.

Case 1: Emotion and Roughness Flattening.
Source (JA): くそ！何やってんだよ！
Official (EN): "Damn it! What the hell are you doing?!"
Modified (EN): "I'm afraid your actions are inappropriate."

	Pol.	Rough.	Emo.	Auth.	Age	Conf.	Dir.
Source	-2	$+1$	$+2$	$+1$	-1	$+2$	$+1$
Official	-1	$+2$	$+2$	$+1$	0	$+1$	$+1$
Modified	$+1$	-1	-1	0	0	0	-1

Diagnostic interpretation: The modification shifts 6 of 7 axes, with the largest deviations on roughness ($|\Delta|$=3), emotion ($|\Delta|$=3), and politeness ($|\Delta|$=2). Directness also drops ($|\Delta|$=2) as the blunt imperative becomes a hedged, formal evaluation, while authority and confidence shift only mildly ($|\Delta|$=1 each). This identifies the stylistic distortion as *emotion flattening with register elevation*, distinct from a single-axis politeness shift or simple lexical substitution.

Case 2: Authority and Confidence Hedging.
Source (JA): 金輪際メガネなんてかけないからな

Official (EN): "I'm never wearing glasses again!"
Modified (EN): "I probably won't wear glasses anymore... I think."

	Pol.	Rough.	Emo.	Auth.	Age	Conf.	Dir.
Source	−1	+1	+1	+2	0	+2	+2
Official	−1	0	+1	+1	0	+2	+1
Modified	0	−1	0	−1	0	−1	−1

Diagnostic interpretation: Primary deviations occur on confidence ($|\Delta|$=3), authority ($|\Delta|$=2), and directness ($|\Delta|$=2). The hedging markers ("probably," "I think") shift epistemic certainty and power direction while increasing processing load, illustrating multi-dimensional stylistic degradation.

Case 3: Politeness and Directness Elevation.
Source (JA): たまには道場の方に寄っていきませんか？
Official (EN): "Maybe you could stop by the dojo?"
Modified (EN): "I would be honored if you might consider visiting our training facility at your convenience."

	Pol.	Rough.	Emo.	Auth.	Age	Conf.	Dir.
Source	0	−1	0	−1	0	−1	+2
Official	0	−1	0	0	0	0	+1
Modified	+2	−1	0	−1	+1	−1	−2

Diagnostic interpretation: The modification shifts politeness ($|\Delta|$=2) and directness ($|\Delta|$=3) while preserving emotion and roughness. This demonstrates axis independence: register and complexity can shift while affective dimensions remain stable.

B Dataset Details

B.1 Subtitle-100

Subtitle-100 is a scorer-validation set consisting of 100 professional JA→EN anime subtitle pairs. All items are taken from works originally produced in Japanese, with Japanese source utterances paired with their official English subtitle translations. Each item is annotated on the seven SDP axes for both the source and target utterances. We use this dataset to validate the LLM-based style scorer and to test whether the scorer's ΔSDP tracks known stylistic deviations.

B.2 Style25

Style25 is a controlled style-perturbation benchmark built in both JA→EN and EN→JA directions. For each direction, we sample 100 seed items from OpenSubtitles (Lison and Tiedemann, 2016). The seed translations are community-produced OpenSubtitles translations.

For each seed translation, Gemma3-27B-IT rewrites the translation into 25 target styles at two intensity levels (Table 9). Together with the unmodified OpenSubtitles translation, this yields up to 51 possible candidates per item before filtering. We apply a COMETKiwi quality gate (≥ 0.80) during generation: candidates below the threshold are regenerated when possible, but those that still fail are discarded. Therefore, fewer than 51 candidates may remain after filtering, with survival rates differing by direction.

We use Style25 to test whether automatic scores respond systematically to controlled style shifts and whether they can identify the unmodified OpenSubtitles translation as the no-perturbation anchor among deliberately style-perturbed variants. This evaluation is diagnostic and should not be interpreted as a measure of translation quality or human preference.

B.2.1 Style25 Prompt Specifications

Each of the 25 styles is defined for both translation directions at two intensity levels. Each level consists of an instruction and a set of examples (3 per level). At prompt-construction time, one example is selected at random from the valid examples to build a 1-shot prompt. The templates are shown below, followed by the full instructions and one representative example per category/intensity.

JA→EN template

```
Translate the following Japanese subtitle
lineinto English.
<DESCRIPTION_ON_STYLE>
Do not add or remove information from the
original. Output ONLY the English translation.

Example:
Japanese: <JAPANESE_EXAMPLE_SENTENCE>
English: <ENGLISH_EXAMPLE_SENTENCE>

Now translate:

Japanese: <SOURCE>
English:
```

EN→JA template

```
以下の英語のセリフを日本語に翻訳してく
ださい。
<DESCRIPTION_ON_STYLE>
原文の情報を追加・削除しないこと。
日本語訳のみを出力してください。

例：
English: <ENGLISH_EXAMPLE_SENTENCE>
日本語:<JAPANESE_EXAMPLE_SENTENCE>
```

#	Category	Style	Instruction (abbrev.)	Example (JA→EN, mild)
Register	01	Textbook	Clean, standard English; educational clarity; no slang.	もう関わりたくない。放っておいてくれ。 → I no longer wish to be involved. Please leave me alone.
	02	Academic	Academic register; nominalization, passive voice.	→ The subject expresses a desire to discontinue involvement and requests non-interference.
	03	Literary	Polished, novelistic prose; aesthetic word choice.	→ A weariness had settled over him—the kind that begs for silence and solitude.
	04	Tabloid	Punchy, sensational headline style.	→ ENOUGH! "Just back off and leave me alone!"
	05	Casual	Conversational tone; contractions, relaxed grammar.	→ Ugh, I'm so over this. Just lemme be, yeah?
	06	SNS	Social-media post; short, informal.	→ done with this lol just leave me alone
	07	Report	Official report; impersonal, third-person.	→ The individual has indicated a preference to discontinue engagement.
Spatiotemporal Variation	08	Regional-1	Regional variety A (Southern / Kansai).	→ I don't wanna deal with this no more. Just leave me be.
	09	Regional-2	Regional variety B (British / Hiroshima).	→ I'd rather not be involved any further, if you don't mind.
	10	Era	1940s period style (film-noir / wartime).	→ I'm through with this racket. Just leave me alone, will ya?
Sociolinguistic Group	11	Business	Corporate register; business jargon.	→ Per our earlier discussion, I'd like to disengage going forward.
	12	Elderly	Gentle, unhurried elderly speech.	→ I've had enough of all this fuss, really. Just let me rest.
	13	Gen-Z	Contemporary youth slang.	→ I literally cannot with this anymore, just leave me alone.
Character Voice	14	Deadpan	Flat, unemotional; short declaratives.	→ I do not want involvement. Leave me alone.
	15	Bitter	Sarcastic, cynical, world-weary.	→ Funny how I keep having to say this—I'm done. Leave me alone.
	16	Brooding	Quiet, introspective; restrained emotion.	→ …I just…I can't do this anymore. Please, just leave me be.
Operational Register	17	Community-register	Expressive, community-associated register.	→ Oh, I am just done with this whole thing. Leave me be, would you?
Multilingual Practice	18	Code-switching	Identity-driven code-switching across languages.	→ I'm done with this, c'est la vie. Just leave me alone.
Pragmatic Strategy	19	Blunt	Direct; no hedging or pleasantries.	→ Done. Leave me alone.
	20	Roundabout	Indirect, roundabout phrasing.	→ I've been thinking, and maybe it might be better if we went our separate ways.
	21	Hedging	Frequent uncertainty markers.	→ I think maybe I probably don't want to be involved anymore…
Stance	22	Assertive	Confident, declarative; no hedging.	→ I'm done. This is over. Leave me alone.
Interpersonal Role	23	Commanding	Authoritative tone; imperative forms.	→ That's enough. Stand down and leave me alone.
	24	Submissive	Timid, deferential toward a superior.	→ Um, I'm sorry, but…I think I'd like to be left alone…
	25	Mentoring	Warm, guiding mentor tone.	→ Listen, I've been where you are. But I need my space now.

Table 9: Style25 taxonomy and style definitions: 25 styles across 9 categories, using the same category names as the main-text figure. Instructions are abbreviated; each style is defined for both translation directions at two intensity levels (mild/strong). All examples share the same source sentence 「もう関わりたくない。放っておいてくれ。」 to illustrate stylistic contrast. The community-register perturbation is a synthetic operational stress test: it does not represent any particular real-world or protected community, and is unrelated to the human-produced OpenSubtitles translations used as unmodified references. Each style is assigned a numeric ID (01–25), matching the numbered points in the PCA figure (Figure 10 and 11).

次の文を翻訳してください：

`English: <SOURCE>`
日本語:

B.2.2 Style25 Style Definitions

The 25 styles are organized into 9 categories. Table 9 lists the full taxonomy together with an abbreviated instruction and one representative example (JA→EN, mild level) for each style. The `<DESCRIPTION_ON_STYLE>` slot in the templates above is filled with the corresponding instruction, and the example sentences are drawn from the examples associated with each style/intensity. For illustration, all examples in Table 9 share the same source sentence 「もう関わりたくない。放っておいてくれ。」 so that stylistic contrasts can be compared directly. Concretely, `<JAPANESE_EXAMPLE_SENTENCE>` and `<ENGLISH_EXAMPLE_SENTENCE>` in the JA→EN template correspond to the source and target sides of the example shown in the rightmost column of Table 9.

B.3 Pro-JE

Pro-JE is a human-preference set of 32 professional JA→EN anime subtitle items, separate from Subtitle-100. Each item contains four candidates: the official human subtitle and outputs from three LLMs (Qwen3-14B-Instruct (Yang et al., 2025), Gemma3-12B-IT (Team et al., 2025), and Llama3.1-8B-Instruct (Grattafiori et al., 2024)).[3] Human preferences are collected using Best–Worst Scaling (BWS). We use Pro-JE for a case study (§6.3) of how SDP dimensions and SFS relate to human preference, not to argue that source-style preservation is generally the preferred outcome.

Candidate generation for Pro-JE. The three LLM candidates for each Pro-JE item were generated by prompting each model to produce a plain, natural English translation of the Japanese source, with no style guidance, so that any stylistic shift reflects the model's default translation behavior rather than an imposed target style. The prompt instructed the model to output only the English translation of the given anime subtitle line. Generation used the chat template of each model with sampling (temperature 0.3, up to 256 new tokens) in bfloat16, and the first non empty output line was taken as the translation after removing surrounding quotes and any reasoning trace markers. This contrasts with Style25 (Appendix B.2), where generation is explicitly style conditioned; in Pro-JE the LLM outputs are unconditioned baselines against which the official human subtitle is compared.

Table 10 summarizes the three datasets used in the experiments that follow.

C Scorer Validation Details

C.1 Annotation Guidelines

All human annotations were produced by in-house bilingual Japanese–English annotators following a shared written guideline. This section summarizes the instructions given to annotators; the per-axis scale descriptions and bilingual anchor examples used during annotation are those in Appendix A (Table 5 and Table 6).

[3] For one Pro-JE item (`pair_089`), Gemma3-12B-IT and Llama3.1-8B-Instruct produced an identical English translation; the duplicate was removed during BWS survey construction, leaving only 3 distinct candidates (Human, Qwen3-14B-Instruct, and the shared Gemma3-12B-IT/Llama3.1-8B-Instruct translation) for this item instead of 4. The resulting 127 (not 128) candidates are used in Appendix F's correlation and quality-gate analyses.

Task and unit of judgment. For each item, annotators independently read the source (Japanese) and target (English) utterances and scored each utterance on all seven SDP axes using the five-point bipolar scale (-2 to $+2$). Japanese and English were scored independently, so that translation-induced style shifts could later be recovered as ΔSDP. Annotators were instructed that there is no correct or incorrect answer and to record their intuitive impression as a number.

Masked, text-only judgment. Speaker identity and surrounding context were hidden. Even when an annotator could infer which character produced a line, they were instructed to ignore character background and judge only the linguistic expression of the given text. For axes whose everyday interpretation depends on the interlocutor or situation, annotators were explicitly told to set aside such context and to score the flat, text-level impression. For example, politeness and authority were judged without assuming a particular speaker–listener relationship, emotion without assuming intimacy from the topic, age from lexical and stylistic markers rather than topic, confidence from epistemic and stylistic markers rather than factual correctness, and directness from the interpretive cost of the utterance rather than from an elaborated imagined context.

Per-axis decision procedure. Each axis was accompanied by a short, stepwise rubric that first removed confounding context, then checked concrete linguistic markers, and finally integrated them into a single score. Representative markers included honorific and register forms for politeness, aggressive or softening expressions for roughness, interjections and emphatic forms for emotion, directive versus deferential framing for authority, age-graded vocabulary and style for age, hedging versus assertive markers for confidence, and lexical difficulty and structural or rhetorical complexity for directness. Annotators were told to assign ± 2 when a property was clearly present and ± 1 when it was mild or when they were uncertain, and to use 0 for neutral text.

Independence and comments. Japanese and English scores were not required to match: divergence between them is the signal of interest. Annotators could optionally leave a free-text comment noting cues they relied on or cases where translation changed the impression, which we use to inform later development. Each annotator worked on an

	Subtitle-100	Style25	Pro-JE
Purpose	scorer validation	automatic style-sensitivity test	supportive human preference analysis
Direction	JA→EN	JA↔EN	JA→EN BWS items
Texts / candidates scored	200 utterances: source and target for each pair	up to 200 × 51 candidates before filtering; fewer after filtering	up to 128 candidates: 32 items × 4 translations (127 scored; one item has only 3, see Appendix B.3)
Judgment units	700 pair-axis judgments: 100 pairs × 7 axes	200 Pres@1 decisions: one per seed item	32 item-level BWS preferences
Human annotation	seven-axis ratings by 3 annotators	none; automatic scoring only	BWS by 4 annotators
Main use	validate the LLM-based SDP scorer	test whether metrics identify the no-perturbation anchor among deliberately style-perturbed candidates	examine whether SDP dimensions relate to human preference

Table 10: Overview of the three datasets. We distinguish items/questions, candidate texts, and judgment units because the datasets differ in structure: Subtitle-100 contains paired source–target utterances, Style25 contains many controlled perturbation candidates per seed item, and Pro-JE contains four translation candidates per item.

independent port, and judgments were saved per item.

C.2 Human Annotation Reliability

For scorer validation, three annotators independently rated Subtitle-100 on all seven SDP axes using a five-point bipolar scale. Annotators were masked to the condition of each item. Because each translation pair requires judgments for both the source and target utterances across seven axes, the annotation cost is substantial: annotating the 100 Subtitle-100 items took 36.5 annotator-hours in total.

Table 11 reports inter-annotator agreement using Krippendorff's α for each axis and language. Agreement varies by axis. Authority is the most reliable axis, whereas age and directness are among the weakest. The mean Krippendorff's α across the seven axes is 0.611 for Japanese and 0.492 for English, with an overall mean of 0.551 across both languages.

This level of agreement is comparable to prior work on pragmatic and social meaning annotation. For example, Pei and Jurgens (2020) reported Krippendorff's $\alpha = 0.548$ over 212 annotated tuples and interpreted it as moderate agreement. Our validation should be interpreted relative to this human agreement level, especially for axes with low inter-annotator agreement.

Axis	α (JA)	α (EN)
Authority	.730	.591
Age	.629	.347
Politeness	.621	.577
Emotion	.616	.530
Confidence	.599	.587
Roughness	.591	.410
Directness	.491	.400

Table 11: Inter-annotator agreement by SDP axis and language. Japanese shows higher agreement than English on all axes. The largest drops for English are observed on age, roughness, and directness.

C.3 Axis-Level LLM Scorer Validation

We further examine the selected Qwen3-14B-Instruct scorer at the axis level. Table 12 reports Spearman's ρ and Kendall's τ for each axis, split by source language (JA) and target language (EN), following the same per-language breakdown used for inter-annotator agreement in Table 11. Agreement is strongest for emotion, confidence, and authority, which are all significant in both languages, whereas age shows the weakest agreement with English-side correlation near zero and directness shows comparatively lower values. This pattern diverges from the human inter-annotator pattern in Table 11, where authority shows the highest agreement and directness the lowest: the LLM scorer tracks human judgments well on emotion and confidence despite those axes involving more subjective distinctions. This suggests that LLM scorer errors are not simply inherited from human-annotation difficulty, but partly reflect axis-specific scorer biases (see the zero-bias analysis in Appendix C.7).

Weak scorer agreement, however, does not necessarily mean that an axis is uninformative. We retain all seven axes because they are part of the predefined SDP schema. The Pro-JE axis-subset

Axis	MAE↓[†]	Spearman ρ		Kendall τ	
		JA	EN	JA	EN
Politeness	.540	.47***	.54***	.43***	.51***
Roughness	.340	.55***	.42***	.52***	.40***
Emotion	.440	.62***	.58***	.57***	.55***
Authority	.395	.47***	.62***	.44***	.59***
Age	.545	.26**	.00	.24**	.00
Confidence	.555	.58***	.59***	.53***	.53***
Directness	.755	.39***	.46***	.34***	.42***

Table 12: Axis-level validation of the selected Qwen3-14B-Instruct scorer against human mean annotations, split by source language (JA) and target language (EN). *$p < .05$, **$p < .01$, ***$p < .001$ (uncorrected for multiple comparisons). [†]MAE is pooled across JA and EN.

Model	±1 Agr.	MAE	Mean ρ
Qwen3-14B-Instruct	**93.9%**	**0.51**	**0.47**
GPT-oss-120B	88.4%	0.69	0.37
Gemma3-9B	87.5%	0.67	0.39

Table 13: LLM scorer comparison. Figures are identical to Table 2 (same 100-item Subtitle-100 set, pair_001–100) and are repeated here for convenience alongside the axis-level breakdown for the selected scorer in Table 12. Mean ρ is the mean per-axis Spearman correlation with human ratings. Qwen3-14B-Instruct achieves the highest ±1 agreement, the lowest MAE, and the highest mean ρ, and is therefore selected as the scorer used throughout the paper.

ablation in Appendix C.9 shows that no small axis subset dominates across objectives (Top-1 selection vs. Kendall's τ), and that scorer agreement alone does not reliably predict which axes carry preference-relevant signal.

C.4 LLM Scorer Comparison

We compared three LLM scorers against human annotations (Table 13). Qwen3-14B-Instruct achieved the highest ±1 agreement, the lowest MAE, and the highest mean per-axis Spearman correlation, and was therefore selected as the scorer used throughout the paper; the axis-level breakdown for this scorer is reported separately in Table 12.

C.5 LLM Scorer Prompts

This section gives the exact prompts for the three LLM-based scorers used in the paper: the SDP scorer (seven axes), our reference-free GEMBA-DA adaptation, and the SA-3 style baseline (RASTA axes).

SDP scorer. Each axis is scored by a separate LLM call (14 calls per item: 7 axes × 2 languages).

Prompt template:

```
Evaluate the {axis_name} of this
{language} sentence on a scale from
-2 to +2.

Sentence: {text}

Scale:
-2: {scale_-2_description}
-1: {scale_-1_description}
0: {scale_0_description}
1: {scale_+1_description}
2: {scale_+2_description}

IMPORTANT: Respond with ONLY a single
number (-2, -1, 0, 1, or 2). Do
not use any tags, explanations, or
thinking process.

Score:
```

Generation parameters:

- Model: Qwen3-14B-Instruct
- Max new tokens: 50
- Temperature: 0.1
- Output format: Label only (no explanation)

GEMBA-DA (reference-free adaptation). As noted in §4, our GEMBA-DA scores follow the Direct Assessment formulation of Kocmi and Federmann (2023) but omit the human-reference field, matching our other reference-free metric (COMETKiwi). Each candidate is scored by a single LLM call.

```
Score the following translation
from {src_lang} to {tgt_lang} on
a continuous scale from 0 to 100,
where:

- 0 means completely wrong,
incomprehensible, or full of errors
- 100 means perfect translation, no
errors, completely natural

Source ({src_lang}): {source}
Translation ({tgt_lang}):
{translation}

Consider accuracy, fluency,
terminology, and style. Provide
only a numeric score.

Score:
```

The returned 0–100 value is normalized to $[0, 1]$. Generation uses Qwen3-14B-Instruct with greedy decoding (max 16 new tokens); the heavy discretization of these greedy scores is analyzed in Appendix D.

Strategy	Zero%	ρ	MAE	Att.%	Rev.%
P0 Baseline	45.7	**.509**	**.600**	32.8	**9.4**
P1 Anti-zero	46.0	.493	.600	32.8	10.9
P2 Few-shot	55.3	.455	.660	42.4	11.0
P3 CoT	39.0	.501	.670	26.6	13.9
P4 Two-stage	41.7	.504	.646	28.1	12.0
P5 Combined	58.7	.431	.638	46.7	10.2

Table 14: Prompt calibration results. No strategy improves all metrics over the baseline. Strategies that reduce zero-rate increase polarity reversals.

SA-3 (RASTA style baseline). SA-3 reuses the three RASTA axes (Havaldar et al., 2025): politeness, intimacy, and formality, each scored on a five-point $[0, 1]$ scale $(0, 0.25, 0.5, 0.75, 1.0)$ by a separate LLM call, mirroring the SDP scorer structure so that the two style measures differ only in their axis inventory.

```
Evaluate the {axis_name} of this
{language} sentence.

{axis_description}

Sentence: {text}

Scale (0 to 1):
0.0: {scale_0.0_description}
0.25: {scale_0.25_description}
0.5: {scale_0.5_description}
0.75: {scale_0.75_description}
1.0: {scale_1.0_description}

IMPORTANT: Respond with ONLY a
single number (0.0, 0.25, 0.5,
0.75, or 1.0). Do not use any tags,
explanations, or thinking process.

Score:
```

Generation uses Qwen3-14B-Instruct (max 50 new tokens, temperature 0.1). SA-3 fidelity is computed from these axis scores using the same MAD-based distance as SFS (Appendix C.8), restricted to the three RASTA axes.

C.6 Prompt Calibration Experiment

Five alternative prompt strategies were tested to determine whether LLM axis attenuation can be mitigated through prompt engineering ($N = 100$ items $\times$ 7 axes $\times$ 2 languages). Results are shown in Table 14.

No strategy improves all metrics over the baseline. Strategies that reduce zero-rate increase polarity reversals. Few-shot examples *increase* zero-rate (55–59%) as anchors function as comparison standards. Baseline achieves highest overall ρ (.509) and lowest reversal rate (9.4%).

Axis	$N_{\neq 0}$	Correct	Atten.	Reversal
politeness	118	53.4%	36.4%	10.2%
roughness	84	40.5%	47.6%	11.9%
emotion	118	72.0%	24.6%	3.4%
authority	74	64.9%	27.0%	8.1%
age	79	27.8%	65.8%	6.3%
confidence	139	61.9%	28.8%	9.4%
directness	146	68.5%	17.1%	14.4%
Total	758	57.8%	32.8%	9.4%

Table 15: Per-axis LLM error decomposition (full 100 items, JA+EN combined). Age has the highest attenuation (65.8%); emotion has the lowest reversal (3.4%). The LLM's primary failure mode is attenuation, not polarity reversal.

Function	d (H)	d (L)	r (H↔L)
MAD_{sim}	**1.01**	**0.84**	**.478**
Norm. Manhattan	1.01	0.84	.478
Norm. Euclidean	1.09	0.71	.435
RMS_{sim}	1.09	0.71	.435
Pearson r	0.78	0.07	.365
Cosine	0.63	0.17	.336
Dot product	-0.23	-0.31	.506
Max deviation	0.85	0.59	.387
Agreement (± 1)	0.72	0.47	.301
Agreement (± 0)	0.41	0.22	.186

Table 16: Distance function comparison. d (H) = Cohen's d for human scoring; d (L) = Cohen's d for LLM scoring; r (H↔L) = correlation between human and LLM composite scores. MAD_{sim} achieves the best joint performance.

C.7 Per-Axis LLM Error Decomposition

LLM scoring errors are decomposed by axis ($N = 100$ items $\times$ 7 axes $\times$ 2 languages) in Table 15. On non-zero human ratings, the LLM preserves polarity in 58% of cases, attenuates to 0 in 33%, and reverses polarity in 9%.

The LLM shows overall zero bias: it outputs 0 on 57% of all ratings versus 46% for humans, an 11-point gap most severe on age (91% vs. 55%) and politeness (54% vs. 35%).

C.8 SFS Distance Function Comparison

Ten distance functions for computing SFS were compared, evaluating each on condition separation (Cohen's d) and Human–LLM agreement (Pearson r). Results are reported in Table 16.

MAD_{sim} achieves the best joint performance. Cosine similarity is invariant to uniform scaling, so it cannot detect "style flattening" where both vectors are proportionally attenuated. Using MAD instead of cosine improved d from 0.63 to 1.01 and r from .336 to .478.

MAD and normalized Manhattan are algebraically identical: $\text{MAD}_{\text{sim}} = 1 - \frac{1}{K}\sum_{i=1}^{K}\frac{|s_i - t_i|}{R}$.

C.9 Axis-Subset Robustness

To test whether the seven-axis SFS is redundant, we exhaustively evaluate all $2^7 - 1 = 127$ non-empty axis subsets on Pro-JE against two objectives: Top-1 selection accuracy and Kendall's τ with human BWS preference. The two objectives favor different axes. The subset that maximizes Top-1 accuracy at k=5 (politeness, roughness, emotion, authority, age) reaches 56.2% but only τ=+0.041, whereas the best subset for τ alone at k=3 (emotion, confidence, directness) reaches τ=+0.195 (above the seven-axis SFS's τ=+0.147), yet is not the Top-1-best subset (k=7: 46.9% Top-1). No subset dominates both objectives simultaneously, and *authority* never appears among the best subsets for either objective, while *emotion* appears in nearly all of them. This shows that human–LLM scorer agreement (Appendix C.3) is not a reliable guide to which axes matter for preference: directness has comparatively strong scorer agreement yet contributes little to Top-1 accuracy, while age has weaker agreement but appears in the best Top-1 subset. We therefore retain the full seven-axis SFS as a fixed, task-agnostic inventory: pruning to any single small subset would optimize one objective at the expense of another, and axis informativeness is objective-dependent rather than reducible to a single ranking.

D Statistical Details

Inter-annotator agreement heatmap. Figure 3 shows Krippendorff's α per axis. This page intentionally includes additional figures for reference.

Condition separation and LLM error profile. Figure 4a shows the SFS distribution by condition (Official vs. Modified) for human and LLM scoring on the full N=100 Subtitle-100 set; both scoring sources clearly separate the two conditions ($d = 0.66$ for $\text{SFS}_{\text{Human}}$, $d = 0.30$ for SFS_{LLM}). Figure 4b shows the per-axis LLM error decomposition (left) and per-axis Human–LLM Spearman ρ (right, N=100 items, JA+EN pooled): age has the highest attenuation rate (65.8%, Table 15) and the weakest axis-level Human–LLM correlation (Table 12), whereas emotion has the lowest reversal rate (3.4%) and among the strongest correlations.

Distance function comparison. Figure 5 visualizes the tradeoff between condition separation and Human–LLM agreement for 10 distance functions.

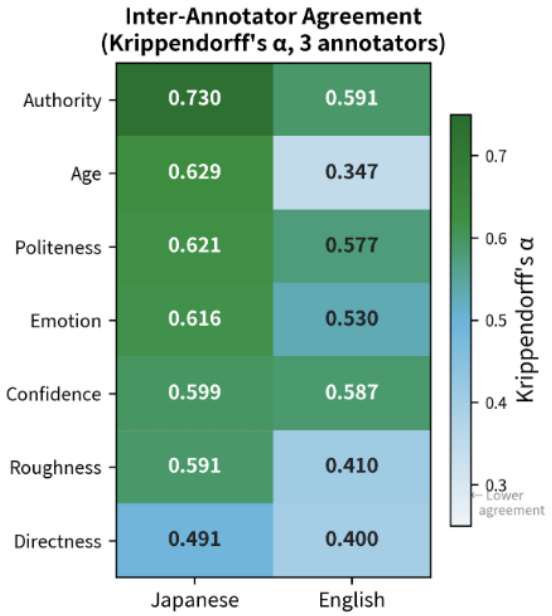
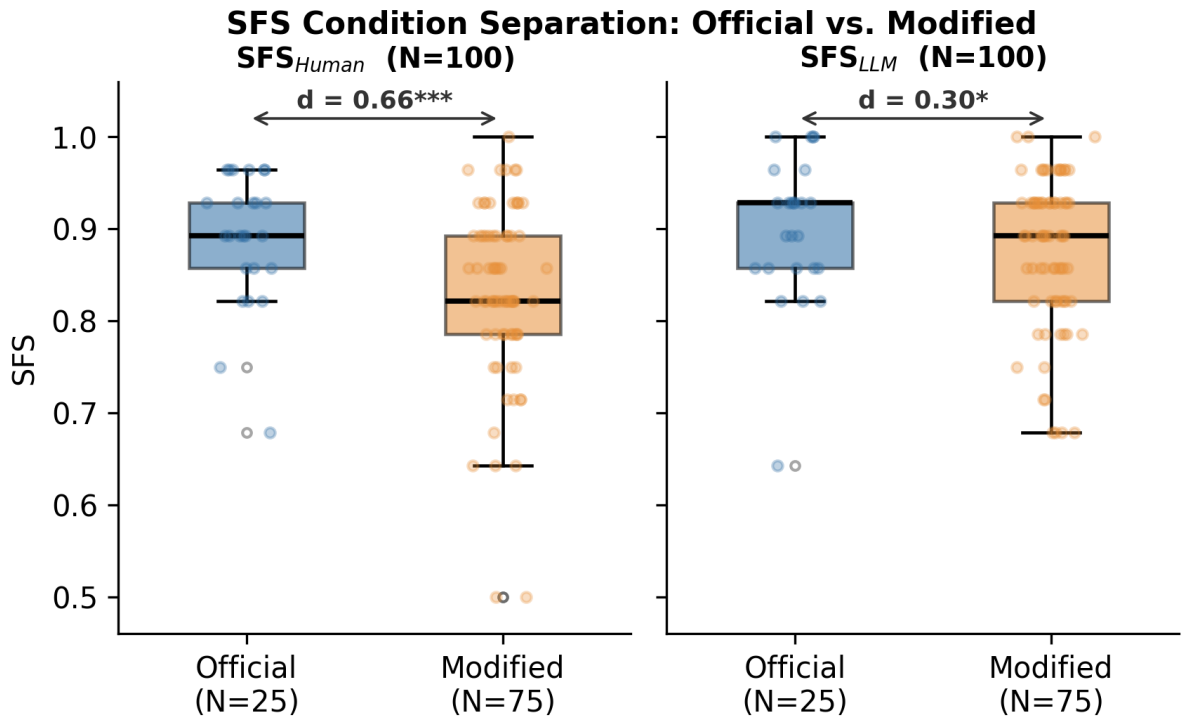


Figure 3: Inter-annotator agreement (Krippendorff's α) per axis.

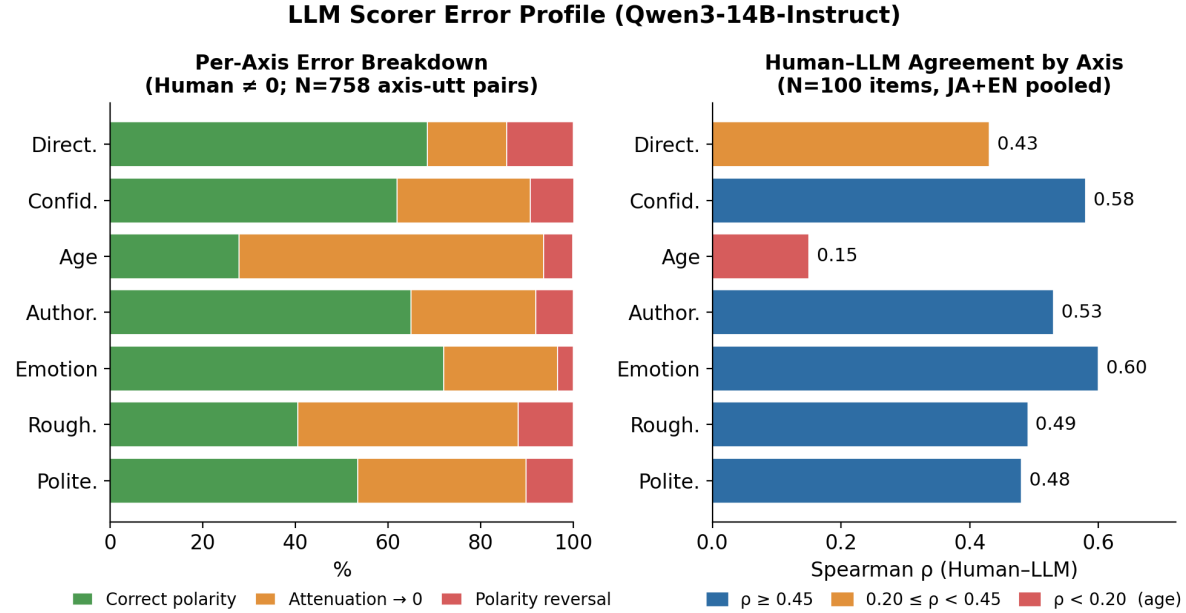
Corpus invariance test. Bootstrap subsampling test (B = 1,000, drawing N = 15 items from 100 JA→EN pairs) illustrates utterance-level design: SFS shows SD = 0.013 (sampling noise only), whereas corpus-level Pearson r fluctuates with SD = 0.12, ranging from 0.08 to 0.75.

GEMBA-DA tie-handling caveat. The GEMBA-DA scores used here come from our reference-free adaptation (§1): an LLM judge is prompted for a single 0–100 score under greedy decoding. In practice the outputs collapse onto a handful of discrete values (6–10 unique values per direction; e.g. 66.8% of JA→EN scores equal exactly 0.95), producing very high anchor–candidate tie rates (54.4% in JA→EN, 33.1% in EN→JA). Such score compression under greedy or low-temperature decoding is a known property of LLM-based DA metrics rather than an artifact of our prompt (Freitag et al., 2024).

Because the anchor is always the first entry in each item's candidate ordering, Pres@1 in Table 3 silently breaks top ties in its favor. Table 17 recomputes Pres@1 under three tie-handling rules: under strict no-tie ranking GEMBA-DA falls to or below chance in both directions, and under fair split-credit it stays below chance in EN→JA while remaining modestly above chance in JA→EN. Thus tie handling explains most, but not all, of GEMBA-



(a) SFS distribution by condition (Official vs. Modified) for human scoring (left) and LLM scoring (right), N=100 items. Both scoring sources clearly separate the two conditions (SFS$_{\text{Human}}$: $d = 0.66$; SFS$_{\text{LLM}}$: $d = 0.30$), confirming that SFS reliably distinguishes perturbed from unperturbed translations regardless of the scoring source.



(b) LLM error profile by axis (left: per-axis error decomposition when Human ≠ 0; right: per-axis Human–LLM Spearman ρ, N=100 items, JA+EN pooled). Age has the highest attenuation rate (65.8%) and the weakest Human–LLM correlation (ρ=0.15); emotion has the lowest reversal rate (3.4%) and among the strongest correlations (ρ=0.60). The dominant error mode is attenuation toward the neutral label rather than polarity reversal.

DA's apparent anchor recovery.

GEMBA-DA's own scores do not support an adequacy-reduction explanation: an item-level comparison of the anchor against its perturbed candidates' mean shows no significant difference in either direction (paired t-test $p = 0.71/0.77$) and no direction asymmetry ($p = 0.79$, Mann–Whitney U). This behavior is expected to generalize: be-

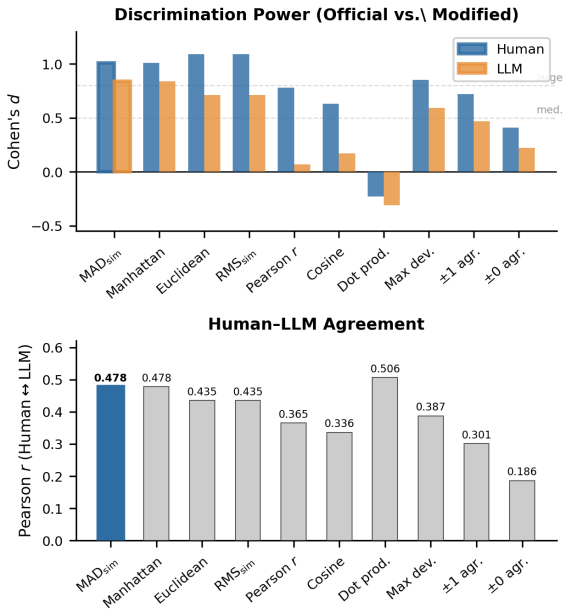


Figure 5: Distance function comparison. MAD_{sim} achieves the best joint performance: the highest discrimination power between Official and Modified conditions (Cohen's d, top) and strong Human–LLM agreement (Pearson r, bottom).

GEMBA-DA tie handling	JA→EN	EN→JA	Type
Standard (Table 3)	74.0	36.0	quality
Strict no-tie	2.0	0.0	quality
Split-credit ties	8.6	2.3	quality

Table 17: GEMBA-DA Pres@1 (%) on Style25 under three tie-handling rules. *Standard* reproduces the ranking used for Table 3, where ties with the anchor are silently broken in its favor; *Strict no-tie* credits the anchor only when its score is the sole maximum; *Split-credit* divides credit evenly among tied top-scoring candidates. Chance levels are 5.7% (JA→EN) and 3.7% (EN→JA). Both corrections fall at or below chance in EN→JA; under split-credit the JA→EN figure remains modestly above chance, so tie handling explains most, but not all, of GEMBA-DA's apparent anchor recovery.

cause the perturbed Style25 candidates preserve meaning and grammaticality (they pass the same COMETKiwi≥ 0.80 gate as the anchor), adequacy-oriented scoring, including reference-based DA and GEMBA-MQM, is by design largely insensitive to style-only changes and would plausibly show comparable compression under greedy decoding.

E Additional Style25 Diagnostics

E.1 Style Shift Magnitude by Category and Intensity

Figure 6 measures the magnitude of the induced shifts. Strong perturbations produce larger mean absolute ΔSDP than mild ones, indicating that ΔSDP is sensitive to the intensity as well as the direction of controlled style changes. The unmodified anchor has the smallest average shift here, a diagnostic property of Style25, not a claim that minimal shift is preferable in professional translation.

E.2 Style Fidelity by Categories

Figure 7 shows mean SFS by category and direction. Style-shift magnitude differs substantially by perturbation type. Lower SFS means a larger shift from the source-style profile, not lower translation quality or human preference.

E.3 Style Adaptation Profile

Figure 8 compares signed ΔSDP profiles across unmodified translations and controlled Style25 perturbations for both translation directions, complementing the professional adaptation analysis in §6.2. The main text finding that professional JA→EN subtitles raise politeness ($+0.24$) and neutralize roughness ($+0.17$) should be read in the context of anime dialogue rather than as evidence of overformalization. Japanese is often described as a language with rich politeness marking, yet many anime utterances involve close social relations, conflict, or strong character voice, and therefore use reduced politeness or highly intimate forms. When such utterances are rendered into idiomatic English subtitles, they frequently map onto a more standard interpersonal register, which appears as a small positive politeness shift on our scale simply because the target expression is closer to a neutral social distance. A parallel effect holds for roughness: very gentle or soft source-side expressions (note the negative source-side roughness, -0.45 in Table 4) tend to be rendered in English with a less marked, more neutral tone, producing a positive roughness shift toward 0 rather than toward crudeness.

These two shifts are therefore not contradictory: both reflect movement toward a standard English interpersonal baseline rather than a uniform increase in formality or aggression. This is consistent with Figure 8, where the unmodified Pro-JE (Human) and Style25 Original rows show smaller and more structured shifts than the controlled LLM-

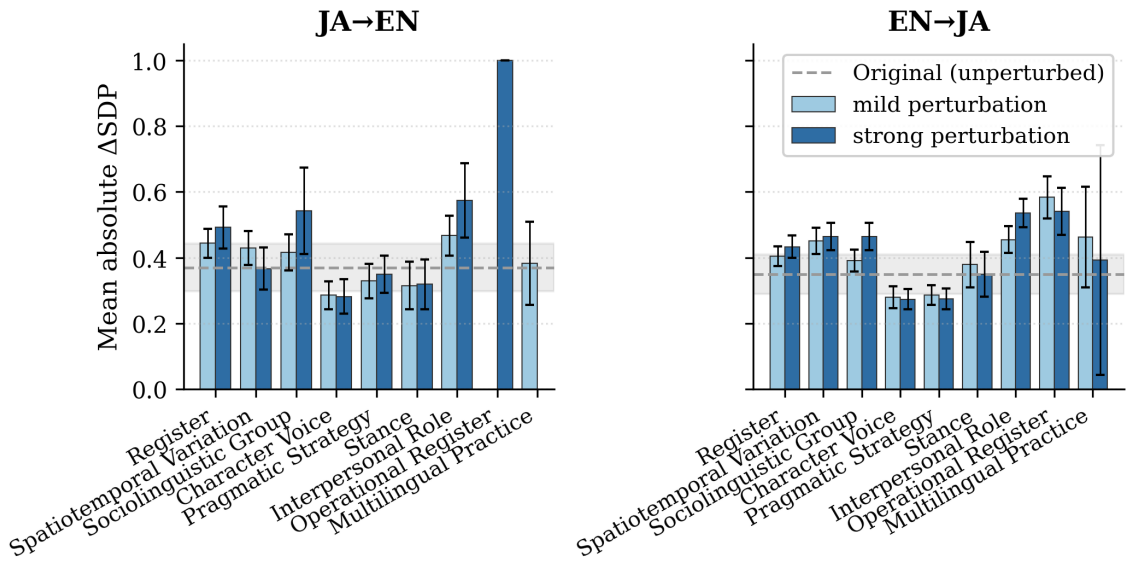


Figure 6: **Style shift magnitude by category and intensity (Style25).** Mean absolute ΔSDP ($\frac{1}{7}\sum_d |\Delta\text{SDP}_d|$) per category, split into mild/strong variants, with the unmodified OpenSubtitles anchor. Strong perturbations generally induce larger shifts than mild ones. Single-style categories, such as community-register and code-switching, are based on fewer samples and should be read with caution.

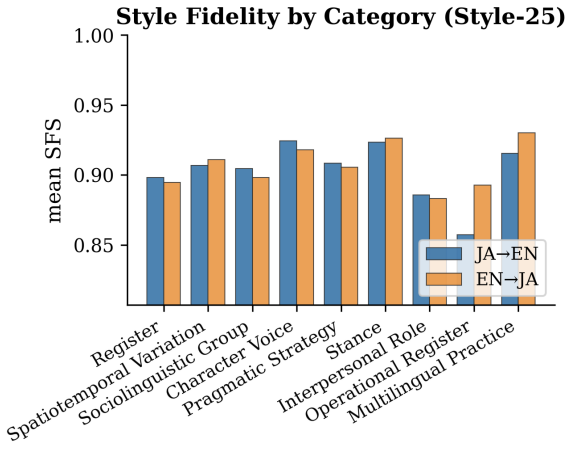


Figure 7: **Style fidelity by category (Style25).** Mean SFS by category and direction. Variation across categories indicates that controlled perturbations induce different magnitudes of style shift. Lower SFS indicates a larger shift from the source-style profile, but does not imply lower translation quality or lower human preference.

perturbation rows. Taken together, the professional sample preserves emotion and directness while normalizing selected interpersonal register cues for the target language, supporting the view that these changes are localization-oriented adaptation rather than style-copying failure.

Category	Dir.	SFS	SA-3	Δ
dialect	JA→EN	78.1	69.9	**+8.2**
locale	JA→EN	60.5	51.2	**+9.3**
character-voice	JA→EN	56.8	50.0	**+6.8**
stance	JA→EN	64.0	73.0	−9.0
register	EN→JA	36.4	51.5	−15.2
speech-style	EN→JA	55.8	42.1	**+13.7**

Table 18: SFS vs. SA-3 anchor recovery by category on Style25 (Pres@1, %). Positive Δ indicates higher Pres@1 for SFS. These six labels are 6 of the 9 categories in the Style25 taxonomy of Table 9, each itself a coarser grouping of the 25 individual Style25 styles; only the categories with the largest $|\Delta|$ are shown here. Correspondence to Table 9: *dialect* = Sociolinguistic Group (Business, Elderly, Gen-Z); *locale* = Spatiotemporal Variation (Regional-1, Regional-2, Era); *character-voice* = Character Voice (Deadpan, Bitter, Brooding); *speech-style* = Pragmatic Strategy (Blunt, Roundabout, Hedging); *stance* = Stance (Assertive); *register* = Register (Textbook, Academic, Literary, Tabloid, Casual, SNS, Report). Interpersonal Role, Operational Register, and Multilingual Practice are not shown here because they did not rank among the largest $|\Delta|$ categories.

E.4 SFS vs. SA-3 by Category

Table 18 breaks down anchor-recovery performance for SFS and the SA-3 baseline by category, restricted to the categories with the largest $|\Delta|$.

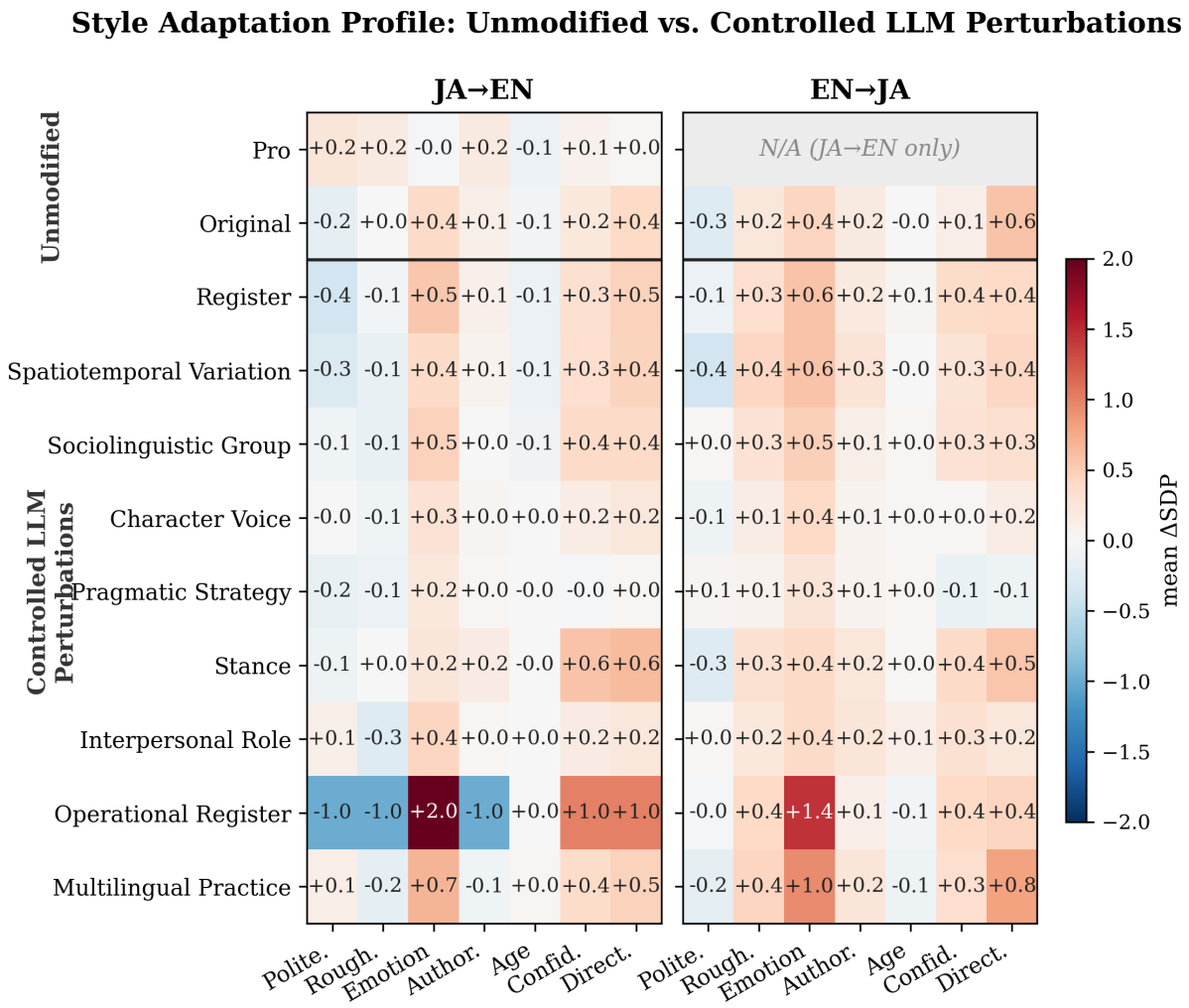


Figure 8: **Style adaptation profile.** Mean signed ΔSDP (translation minus source) for JA→EN (left) and EN→JA (right) on a shared colour scale. Rows are grouped into *Unmodified* translations (Pro-JE professional Human, available for JA→EN only, and Style25 Original) and *Controlled LLM Perturbations* (Style25 categories). The unmodified rows show smaller and more structured shifts than the controlled perturbation rows, while LLM perturbations induce larger changes especially in emotion, confidence, and directness. This supports the view that style changes should be analyzed as signed adaptation profiles rather than treated simply as copying or failure to copy. The community-register condition is an operational stress test and does not represent a real community.

F Full Per-Metric Correlations and Quality-Gate Analysis on Pro-JE

This appendix reports the full per-metric correlations and quality-gate analysis on Pro-JE referenced in §6.3. All results use Pro-JE only (32 items; up to 128 candidates) and are reported as case-study observations (§4.3). Significance markers follow $^{\dagger}p < .10$, $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$ (two-sided).

Top-1 accuracy with fractional credit. For Top-1 accuracy, metric ties receive fractional credit: if the metric's top set $\mathcal{M}_{\text{top}}$ contains n tied candidates, the credit is

$$\frac{|\mathcal{M}_{\text{top}} \cap \mathcal{B}_{\text{top}}|}{n}, \tag{4}$$

where $\mathcal{B}_{\text{top}}$ is the set of BWS-best candidates, and BWS ties are treated as an equally best set. This is the Top-1 definition used both in the main text (§4.3) and throughout this appendix.

F.1 Per-Metric Correlation with Human BWS

Table 19 reports, for each candidate metric, Kendall's τ against human BWS scores and Top-1 accuracy (the fraction of items for which the metric's top candidate matches the BWS-best candidate). Chance-level Top-1 accuracy is 0.250 (one of four candidates per item).

Metric	Type	τ	Top-1	Note
COMETKiwi	Adequacy	-0.008	0.438	Uncorrelated with BWS on Pro-JE
MetricX-24	Quality	$+0.038$	0.469	Weak overall
GEMBA-DA	Quality	$+0.160^{*}$	0.380	Strongest τ among all metrics; ties on 84% of items lower its Top-1 (see note below)
SFS (7-axis)	Style	$+0.147^{*}$	0.404	Style signal, largely independent of quality
SA-3 (RASTA axes)	Style	$+0.008$	0.307	Weak on Pro-JE

Table 19: Per-metric correlation with human BWS preference on Pro-JE. Chance-level Top-1 accuracy is 0.250. Metric ties receive fractional credit (credit $= |\mathcal{M}_{\text{top}} \cap \mathcal{B}_{\text{top}}|/n$, where n is the number of tied top candidates); BWS ties are treated as an equally-best set (§4.3).

GEMBA-DA attains the strongest rank correlation with human preference on Pro-JE ($\tau = +0.160^{*}$) but a comparatively low Top-1 accuracy (0.380): on 84% of Pro-JE items, GEMBA-DA assigns all four candidates the same score (0.95), so its top-ranked set is frequently a four-way tie. Under fractional-credit scoring this depresses Top-1 accuracy even though the metric's ordinal correlation with BWS is the strongest observed. This illustrates that τ and Top-1 accuracy capture different aspects of alignment with human preference and should be interpreted jointly rather than in isolation. COMETKiwi, an adequacy-oriented metric, is uncorrelated with BWS on Pro-JE ($\tau \approx -0.008$), consistent with the claim in §6.3 that style fidelity (SFS) provides preference-relevant signal that is complementary to, rather than redundant with, quality-oriented metrics. Applying Holm–Bonferroni correction across these five tests ($m = 5$) raises the two nominally significant p-values (GEMBA-DA $p = .030$; SFS $p = .032$) above the corrected threshold ($p_{\text{Holm}} = .148$ for both), so neither survives family-wise correction; we report this openly rather than treat the uncorrected markers as conclusive.

F.2 Quality-Gate Sensitivity

Table 20 shows how the SFS–BWS and SA-3–BWS associations change when candidates are restricted to those passing increasingly strict quality thresholds (COMETKiwi, MetricX-24, and/or GEMBA-DA), starting from all 127 scored Pro-JE candidates.

Quality gating consistently strengthens, and never reverses, the SFS–BWS association: SFS τ ranges from $+0.116$ to $+0.225^{*}$ across gates, compared to $+0.147^{*}$ unfiltered, and SFS Top-1 accuracy rises from 0.404 (unfiltered) to as high as 0.656 under the combined COMET+GEMBA gate. In contrast, SA-3 remains uncorrelated or weakly negatively correlated with BWS under every gate (τ between -0.041 and $+0.008$). This pattern supports interpreting semantic adequacy and style fidelity as jointly necessary, rather than substitutable, signals for human preference on Pro-JE: once low-adequacy candidates are removed, the residual variance in human preference is better explained by style fidelity (SFS) than by the narrower RASTA-derived axis set (SA-3). Treating all twelve SFS/SA-3 tests across the six gate conditions as one family, Holm–Bonferroni correction ($m = 12$) leaves none significant at $\alpha = .05$ (smallest raw $p = .032$ against a corrected threshold of .0042); the gate-wise pattern above is therefore best read as a consistent descriptive trend rather than a family-wise-significant effect.

F.3 Axis-Level Association with Human Preference

Table 21 and Table 22 report Kendall's τ between each individual style axis and human BWS scores on Pro-JE, for the SDP seven-axis set and the SA-3 baseline, respectively.

On Pro-JE, no single SA-3 axis reaches significance, and among the seven SDP axes only emotion is significant ($\tau = +0.153^{*}$), with confidence showing a marginal positive trend ($\tau = +0.139^{\dagger}$). This axis-level breakdown is consistent with §6.2's finding that professional translators leave emotion and directness nearly unchanged on average: the present result further suggests that, when translations do vary in emotional expression, human raters on Pro-JE prefer the more emotionally expressive candidate.

F.4 Confound Check Between Axis Values and Quality Gate

Because §F.2 restricts candidates by quality-metric thresholds, an axis–preference association could

Quality gate	Pass N	SFS τ	SFS Top-1	SA-3 τ	SA-3 Top-1
(none)	127	$+0.147^{*}$	0.404	$+0.008$	0.307
COMET ≥ 0.8	62	$+0.161$	0.575	-0.010	0.383
MetricX ≥ 0.895	63	$+0.116$	0.533	$+0.006$	0.479
GEMBA ≥ 0.950	95	$+0.143^{\dagger}$	0.512	-0.019	0.366
COMET + MetricX	50	$+0.129$	0.559	-0.041	0.422
COMET + GEMBA	51	$+\mathbf{0.225}^{*}$	**0.656**	-0.028	0.354

Table 20: Effect of quality gating on the SFS–BWS and SA-3–BWS associations on Pro-JE. Pass N is the number of Pro-JE candidates retained after applying the quality threshold(s). Bold marks configurations whose SFS association exceeds the unfiltered (none) baseline on both τ and Top-1.

SDP axis	τ	Note
Politeness	-0.051	n.s.
Roughness	$+0.002$	n.s.
Emotion	$+0.153^{*}$	Strongest single-axis association on Pro-JE; more emotional candidates are preferred
Authority	$+0.041$	n.s.
Age	-0.120	Negative trend, n.s.
Confidence	$+0.139^{\dagger}$	Weak positive trend (marginal)
Directness	$+0.108$	n.s.

Table 21: Axis-level Kendall's τ between SDP axes and human BWS on Pro-JE.

SA-3 axis	τ	Note
Politeness	$+0.071$	n.s.
Intimacy	$+0.004$	n.s.
Formality	$+0.032$	n.s.

Table 22: Axis-level Kendall's τ between SA-3 axes and human BWS on Pro-JE.

Family	Axis	η^2
SA-3	Formality	0.000
SA-3	Intimacy	0.011
SA-3	Politeness	0.009
SDP	Politeness	0.035
SDP	Emotion	0.025
SDP	Directness	0.044
SDP	Age	0.016
SDP	Roughness	0.003
SDP	Authority	0.009

Table 23: Confound check: η^2 between each style axis and quality-gate pass/fail status on Pro-JE. All values are small (≤ 0.044), indicating no substantial confound between axis values and the quality gate in the Pro-JE candidate set.

in principle be inflated if axis values happen to correlate with the pass/fail quality-gate status itself. Table 23 reports η^2 between each axis and the (COMET + GEMBA) quality-gate pass/fail indicator on Pro-JE.

All η^2 values on Pro-JE are small (≤ 0.044), indicating that, unlike the Style25-derived controlled perturbation sets, the Pro-JE candidate pool (official subtitle plus three LLM outputs) does not exhibit a systematic relationship between axis values and quality-gate status. Pro-JE is therefore an appropriate baseline for interpreting the axis-level and quality-gate associations reported above without confound-driven inflation.

F.5 Pooled versus Within-Group Decomposition

As a further robustness check, Table 24 decomposes the pooled Kendall's τ between each axis and BWS into a within-group component (computed after conditioning on the COMET+GEMBA quality-gate group) and a between-group inflation component (pooled τ minus within-group τ), for the axes with the largest pooled associations on Pro-JE. Within τ and Residual τ are two distinct within-group estimators rather than the same quantity under different names: Within τ is the size-weighted average of Kendall's τ computed separately inside each quality-gate group, whereas Residual τ is a single pooled τ computed after group-mean-centering both the axis value and BWS within each group. The two can diverge because Kendall's τ is not linear, so averaging per-group correlations is not equivalent to correlating pooled, group-centered residuals.

On Pro-JE, most of the pooled association for SA-3 formality and SDP emotion survives within-group conditioning (residual τ of $+0.053$ and $+0.065$, respectively), whereas SDP directness shows a larger inflation component ($+0.105$) relative to its within-group residual ($+0.020$), suggesting that its pooled association is driven more by between-group differences than by within-group rank agreement. SDP

Axis	Pooled τ	Within τ	Inflation	Residual τ
SA-3 Formality	$+0.085$	$+0.073$	$+0.012$	$+0.053$
SDP Emotion	$+0.123$	$+0.068$	$+0.055$	$+0.065$
SDP Directness	$+0.125^{\dagger}$	$+0.020$	$+0.105$	$+0.026$
SDP Confidence	$+0.105$	$+0.136$	-0.031	$+0.110^{\dagger}$

Table 24: Pooled vs. within-group (COMET+GEMBA quality-gate group) Kendall's τ on Pro-JE. Inflation is the pooled-minus-within difference, i.e., the portion of the pooled association attributable to between-group mean differences rather than within-group rank agreement.

confidence shows negative inflation, i.e., its within-group association ($+0.136$) is if anything stronger than the pooled association. Together with the confound check in §F.4, these decompositions indicate that Pro-JE's candidate design does not introduce the kind of confound structure observed in the Style25-derived comparison sets, supporting its use as the primary basis for the case-study human-preference claims in §6.3.

G Direction-Specific Style Planes for Style25

To complement the pooled Style25 visualization in Figure 2, we plot the same interpersonal–expressive style plane separately for JA→EN and EN→JA. The two composite axes are defined as in §6.1: *interpersonal force* $=$ mean(Δdir, Δaut, $-\Delta$pol) and *expressive intensity* $=$ mean(Δemo, Δrou, Δcon).

As shown in Figure 9, Style25 categories occupy different regions in both directions, supporting the pooled analysis. The unmodified anchor is not exactly at the origin in either direction, indicating that naturally occurring translations may already introduce cross-lingual shifts. These plots serve only as diagnostic evidence that ΔSDP captures systematic style-shift patterns, not as quality or preference judgments.

H Data-Driven Projection of Style25 Category Shifts

The two-dimensional view in Figure 2 and Figure 9 uses a hand-chosen projection (§6.1). To check that the observed category separation is not an artifact of this specific projection, we also apply principal component analysis (PCA) to the full seven-axis ΔSDP vectors of all Style25 variants, pooling both translation directions.

Figure 10 shows the result. The first two components explain 75.5% and 13.2% of the variance (88.7% cumulative), indicating that most category-level variation in the seven-axis space is captured by a two-dimensional structure. The point is that the same categories that are separated under the hand-chosen projection remain separated under this data-driven projection, indicating that the separation in Figure 2 is not only for the particular axis combination we chose for visualization. The full seven-axis ΔSDP representation can be examined through other axis pairs or combinations depending on the stylistic contrast of interest, and the PCA plane used here is only one such view that happens to capture most of the variance (PC1 + PC2 $=$ 88.7%).

The figure also plots the individual base-style means (light points), making within-category and cross-style variation directly visible. To make this within-category structure easier to read, Figure 11 re-plots the same base-style means on an identical PCA scale, split into nine per-category panels. As this shows, categories differ in internal spread: multi-style categories such as Register (n=7), Pragmatic Strategy (n=3) and Interpersonal Role (n=3) contain styles that spread widely, sometimes in opposing directions, whereas single-style categories (Stance, Community-register perturbation, and Multilingual Practice, each n=1) are represented by a single point and have no within-category variation by construction. We take this to suggest that our categories are a convenient grouping, and that styles placed in the same category may respond along different axes.

Two aspects of the seven-axis representation deserve comment; the categories themselves are used only as a convenient grouping for display. First, the separation in Figure 2 is not an artifact of our hand-chosen projection. Under a data-driven PCA projection, the categories still separate from the unmodified anchor and from one another, lying away from the origin rather than collapsing toward it, so the category-level means do not result from large internal cancellation. Second, a low-dimensional view, whether hand-chosen or data-driven, is a reading aid rather than the representation our diagnostics operate on, and the high variance captured by PC1 does not imply that fewer axes suffice. PCA maximizes explained variance, not diagnostic interpretability: a principal component mixes axes and cannot by itself indicate *which* stylistic property changed, whereas SDP attributes a shift to named, independently meaningful axes (as in the per-axis Pro-JE analysis of §6.2). The within-category spread in Figure 11 further shows that

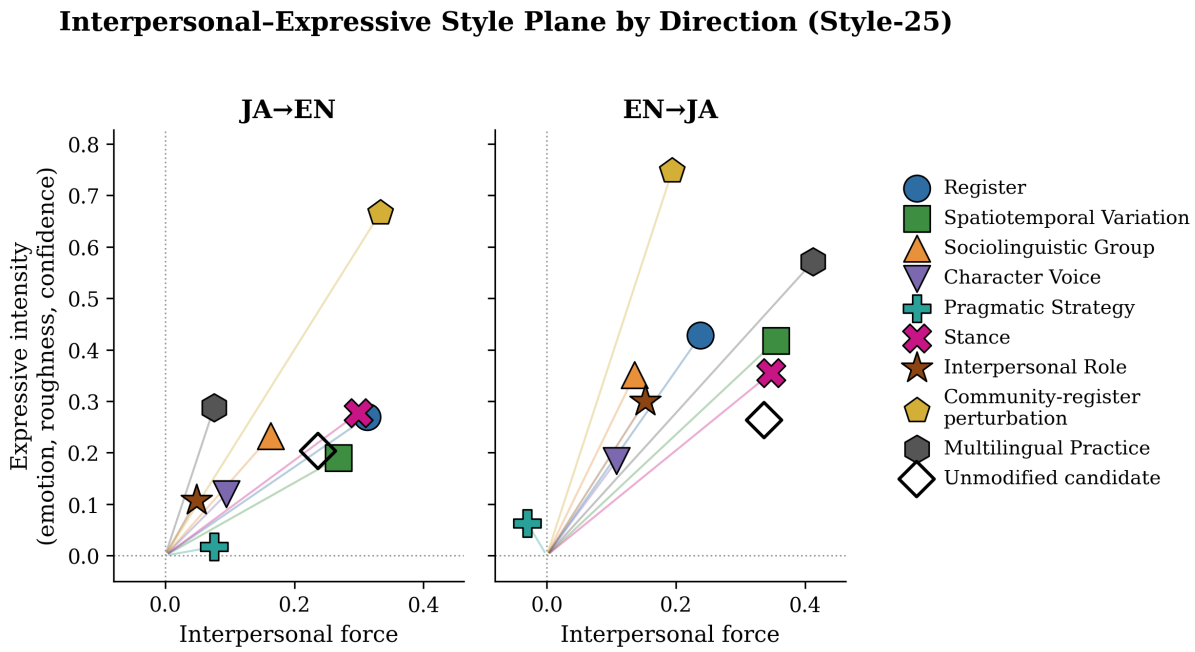


Figure 9: **Interpersonal–expressive style plane by direction (Style25).** Each point is a Style25 category placed by its mean signed ΔSDP separately for JA→EN and EN→JA. The white diamond is the unmodified OpenSubtitles anchor. As in the pooled view (Figure 2), categories occupy distinct regions in both directions, confirming that the category-specific pattern is not an artifact of pooling across translation directions.

styles sharing similar PC1 and PC2 coordinates can differ substantially on the remaining axes, so individual styles are not fully recovered by any two dimensions. We therefore report the full seven-axis SDP profile and treat Figure 2 and Figure 10 as visual summaries only.

I Case Studies

Table 25 illustrates how ΔSDP diagnoses human preference at the item level. Here COMETKiwi scores Qwen3-14B-Instruct highest, but human BWS strongly prefers the official subtitle. The official subtitle is not a style-identical copy; it slightly reduces roughness and directness, yet it preserves the source's emotional force and informal character voice through its hesitation and expressive phrasing. The LLM outputs instead neutralize the utterance by reducing roughness, increasing confidence, and sometimes reducing directness or emotion. ΔSDP thus identifies which stylistic dimensions shift, explaining preference patterns that scalar quality scores miss.

Beyond this single item, Table 26 reports candidate-level aggregates over all $N = 32$ Pro-JE items. The Human (Official) translation attains the highest group-mean BWS ($+0.358$) and is chosen best most often (50.8%), while Llama3.1-8B is the least preferred (-0.345) and most often chosen worst (44.8%). The same axes that distinguish the candidates in the item-level example also separate them in aggregate: the Human translation shows the smallest mean emotion shift ($|\Delta\text{emo}| = 0.22$), whereas Llama shows the largest mean directness shift ($|\Delta\text{dir}| = 1.07$). This pattern is consistent with the interpretation in §6.3: preferred translations tend to preserve the source's emotional force, while less preferred ones drift most on directness. We report these as descriptive case-study observations, not confirmatory findings.

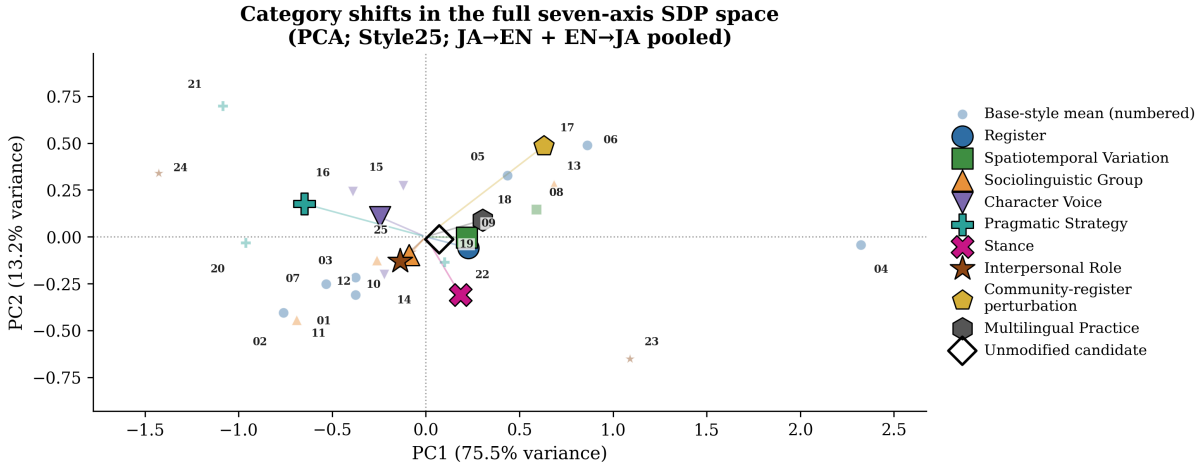


Figure 10: **Category shifts in the full seven-axis SDP space (PCA; Style25, JA↔EN pooled).** PCA on the seven-axis ΔSDP vectors. Large markers are the nine category means; light points are individual base-style means, showing within-category variation. Numbers (01–25) correspond to the style IDs in Table 9. The white diamond is the unmodified OpenSubtitles anchor. The first two components explain 75.5% and 13.2% of the variance.

Candidate	Translation	BWS	SFS	COMET	Major ΔSDP
Human	W-wow... Talk about a supersized ambition...	1.00	0.93	0.79	Rou −1, Dir −1
Gemma3-12B	That's an incredibly ambitious goal.	−0.75	0.89	0.68	Pol +1, Rou −1, Conf +1
Qwen3-14B-Instruct	Wow, the scale of the ambition is huge.	0.00	0.86	0.83	Pol +1, Rou −1, Conf +1, Dir −1
Llama3.1-8B	That's a pretty huge ambition.	−0.25	0.82	0.76	Pol +1, Rou −1, Emo −1, Conf +1, Dir −1

Table 25: **Case study.** ΔSDP diagnoses item-level differences between human preference and scalar quality scores. COMETKiwi ranks Qwen3-14B-Instruct highest, whereas human BWS prefers the official subtitle. The preferred translation is not a perfect style copy, but better preserves the source's emotional and informal character voice. Source (JA):すげえ 野望のサイズがデカい(pair_070).

Candidate	BWS Score	Best (%)	Worst (%)	mean |Δemo|	mean |Δdir|
Human (Official)	**+0.358**	50.8	15.0	**0.22**	0.63
Qwen3-14B	+0.050	22.5	17.5	0.41	0.72
Gemma3-12B	−0.075	16.7	24.2	0.41	0.78
Llama3.1-8B	−0.345	10.3	44.8	0.42	**1.07**

Table 26: Candidate-level human preference (BWS) and mean absolute style shift on Pro-JE ($N = 32$). The most-preferred Human translation shows the smallest emotion shift, while the least-preferred Llama shows the largest directness shift.

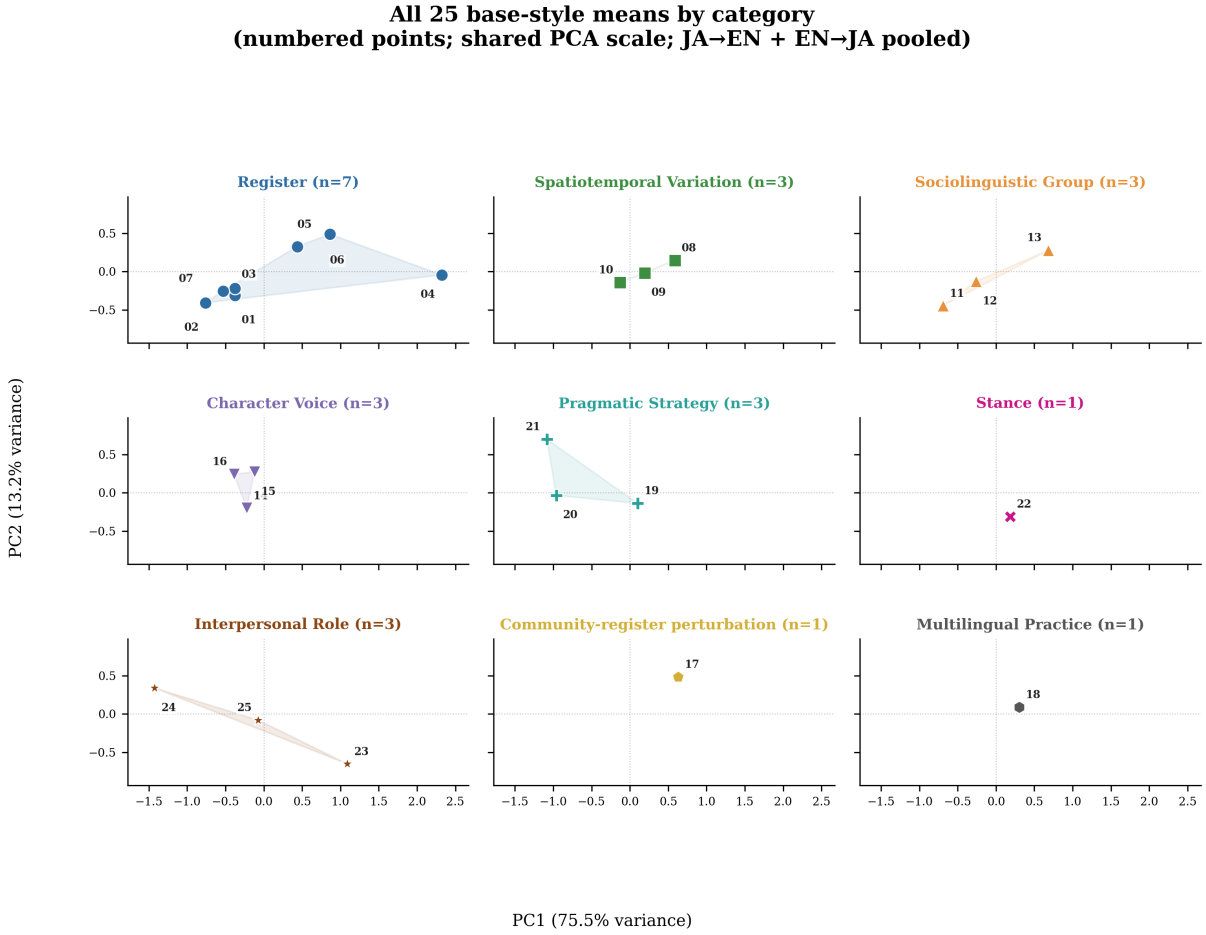


Figure 11: **Per-category view of the 25 base-style means in the full seven-axis SDP space (PCA; Style25, JA↔EN pooled).** The same base-style means as in Figure 10, plotted on an identical PCA scale but split into nine per-category panels. Numbers (01–25) correspond to the style IDs in Table 9. Shaded hulls indicate the spread within each category, making within-category and cross-style variation directly visible; no hull is drawn for categories with a single style ($n = 1$)