Does Translation Preserve Character Personality? Quantifying Perceived Character-Personality Consistency Across Languages

Xinger Fu[1] **Masaaki Nagata**[2] **Chenhui Chu**[1,3]
[1]Kyoto University [2]NTT Communication Science Laboratories, NTT, Inc. [3]NII LLMC
xinger@nlp.ist.i.kyoto-u.ac.jp masaaki.nagata@ntt.com
chu@i.kyoto-u.ac.jp

Abstract

Translation conveys meaning across languages, but does it preserve *the personality a character conveys*? We introduce an evaluation framework for quantifying how well translation systems maintain perceived character personality, the personality impression a reader forms from a character's dialogue, described along the five Big Five dimensions; we use these dimensions as descriptive labels for the impression, not as a psychometric measurement of the character. Our framework applies personality predictors to both source and translated utterances, measuring cross-lingual consistency via Spearman rank correlation at the speaker level. We study entertainment subtitle translation, where every utterance is produced by a named character, and compare evaluation conditions that include, shuffle, or withhold the character names. We evaluate five MT systems, including dedicated MT APIs (Google Translate, DeepL) and LLMs (GPT-4o, Qwen3-32B, Aya-expanse-32B), and additionally report human translation as a reference condition on GAME OF THRONES (English→12 target languages, 11 characters) and ONE PIECE (Japanese→English, 15 characters). Under our evaluation protocol, measured personality consistency varies across systems and language pairs, and high BLEU/COMET translation quality does not reliably imply high personality consistency.

1 Introduction

Translation has long been studied as a meaning-preservation task: does the translation convey the same *content* as the source? Automatic metrics such as BLEU (Papineni et al., 2002), COMET (Rei et al., 2020), and BERTScore (Zhang et al., 2020) measure this content fidelity. Yet natural dialogue carries far more than propositional content: it encodes the speaker's *personality*, *social identity*, and *interpersonal style*. When a sarcastic speaker's lines are translated into bland, neutral prose, the meaning may survive but the *person* is lost.

Despite growing interest in speaker-aware and style-aware machine translation (MT) (Mirkin et al., 2015; Michel and Neubig, 2018), there is no established framework for *quantitatively measuring* how well translation maintains perceived character personality. Prior work has shown that author traits such as gender, age, and personality traits leave detectable signals in text (Argamon et al., 2009; Schwartz et al., 2013), and that these signals can be disrupted by translation (Rabinovich et al., 2017). However, systematic evaluation across multiple language pairs, translation systems, and personality traits remains lacking.

In this paper, we address this gap by proposing a **cross-lingual personality consistency evaluation framework** based on the Big Five personality model, which covers OPENNESS (O), CONSCIENTIOUSNESS (C), EXTRAVERSION (E), AGREEABLENESS (A), and NEUROTICISM (N) and is the dominant framework for personality research due to its cross-cultural validity (Goldberg, 1990; John and Srivastava, 1999; McCrae and Costa, Jr., 1997). We adopt LLM-based personality prediction as our measurement tool: unlike fine-tuned encoder models that require per-language training data (Fu et al., 2024), a single multilingual LLM can serve as a predictor across all languages with consistent scoring via structured output. We apply this predictor to both source and translated utterances, then measure how consistently each character's predicted trait profile carries over into the translation. We position this work as an exploratory study of *perceived, LLM-inferred character-personality consistency* in fictional dialogue: the scores measure the personality impression that a dialogue-reading predictor forms and how stable that impression is across languages, not latent psychometric personality (§3).

Our framework targets entertainment content translation, localization, and transcreation, where

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 4338–4358
November 6-10, 2026 ©2026 Association for Computational Linguistics

professional translators routinely work with character context such as names and biographies. Viewers likewise encounter dialogue alongside character identity. Our main tables report the **Named** condition (the predictor sees the original character names), and we compare it against Shuffled, Anonymized, and name-only conditions to separate what the dialogue contributes from what the name contributes (§5.1).

Our empirical study spans two dialogue corpora: GAME OF THRONES (English→12 target languages, 11 characters) and ONE PIECE (Japanese→English, 15 characters), five MT systems spanning dedicated MT APIs and LLMs, and a human-translation reference condition. Our experiments yield two findings:

1. Personality consistency varies across systems and language pairs, with no single system consistently dominating.

2. Reference-based BLEU/COMET translation quality does not reliably track personality consistency.

The contributions of this work are two-fold:

1. A cross-lingual personality consistency evaluation framework with metrics at both speaker and personality trait levels.

2. A multilingual empirical study spanning five MT systems, human translation, 13 language pairs across two corpora.

2 Related Work

Personality in Dialogue. Computational personality prediction has a rich history, from early feature-based approaches (Mairesse et al., 2007; Schwartz et al., 2013) to modern neural methods (Jiang et al., 2020; Yang et al., 2023a). Historically, much work has focused on English, using self-reported questionnaires or crowdsourced annotations as ground truth. Recent work has explored personality assessment and induction in LLMs: Jiang et al. (2023) propose a psychometric framework for evaluating and inducing personality in pre-trained language models, while Serapio-García et al. (2025) study the reliability, validity, and steerability of personality traits expressed by LLMs. Related dialogue-generation studies focus on persona consistency, empathy, or personality-conditioned generation rather than translation, including persona-based conversation modeling (Li et al., 2016), empathetic dialogue modeling (Rashkin et al., 2019), and personality-aware empathetic stylization in StyEmp (Fu et al., 2024). More recent zero-shot approaches leverage chain-of-thought reasoning over psychological questionnaire items to enhance LLM-based personality detection without task-specific training (Yang et al., 2023b). Psychometrically grounded benchmarks such as TRAIT (Lee et al., 2025) have been proposed to rigorously evaluate the validity and reliability of LLM personality measurements. Beyond prediction, personality-*conditioned* generation has advanced from prompting to training-based approaches. Li et al. (2025) demonstrate that supervised fine-tuning and preference optimization on human-grounded personality dialogue data produce more robust personality shaping than prompting alone. Sun et al. (2025) show that personality can be continuously modulated via model merging, constructing "personality vectors" from the weight difference between trait-finetuned and base models. In the cross-lingual dimension. Amidei et al. (2025) find that language switching itself induces personality trait variation in multilingual LLMs, suggesting that cross-lingual personality differences may partly stem from language-specific model behavior rather than translation quality alone, a concern directly relevant to our evaluation framework.

Author Traits and Translation. Rabinovich et al. (2017) showed that author demographic traits (gender, age) can be disrupted by translation, and proposed trait-preserving MT models. Mirkin et al. (2015) introduced the concept of "personalized machine translation" adapting output to author characteristics. Michel and Neubig (2018) proposed extreme adaptation to personalize neural MT output for individual users. However, these works focus on surface-level stylistic features rather than structured personality models with quantitative evaluation.

Speaker-Aware Translation. In translation studies, preserving character voice is a long-recognized challenge (Nida, 1964; Munday et al., 2022). Recent work on document-level translation has leveraged broader discourse context (Wang et al., 2023), but the preservation of perceived character personality remains unevaluated. Personality-aware prompting for LLM translation conditioning on Big Five traits can be a natural extension. In this work we focus on *measuring* whether translation systems

maintain perceived character personality, and leave systematic personality-conditioned translation as future work.

Multilingual Personality Prediction. While the Big Five model has been validated across cultures (McCrae and Costa, Jr., 1997; Schmitt et al., 2007), *automated* personality prediction in multilingual settings remains underexplored. Sun et al. (2024) introduced a Chinese personality evidence dataset, and the RealPersonaChat corpus (Yamashita et al., 2023) provides Japanese personality-annotated dialogues. These resources have primarily supported language-specific personality modeling, typically with encoder-based approaches rather than translation-focused multilingual evaluation.

3 Methodology

3.1 Problem Formulation

Figure 1 illustrates our evaluation framework. Let S denote the set of speakers. For each speaker $s_i \in S$, let $U^{\text{src}}_{s_i} = \{u^{\text{src}}_{i,j}\}_{j=1}^{n_i}$ denote the source-language utterance set. Here, $n_i = |U^{\text{src}}_{s_i}|$ is the number of utterances for speaker s_i, and j indexes utterances within that speaker's set. A translation system T is applied utterance-wise to obtain $U^{\text{tgt}}_{s_i} = \{u^{\text{tgt}}_{i,j}\}_{j=1}^{n_i}$, where $u^{\text{tgt}}_{i,j} = T(u^{\text{src}}_{i,j})$. A personality predictor $f : \mathcal{U}^* \to \mathbb{R}^5$ maps each speaker-level utterance set to a Big Five personality vector $\mathbf{p} = (p_O, p_C, p_E, p_A, p_N) \in [0,1]^5$, where each component is a trait score.

For each speaker $s_i \in S$, we compute:

$$\mathbf{p}^{\text{src}}_{s_i} = f(U^{\text{src}}_{s_i}) \quad (1)$$

$$\mathbf{p}^{\text{tgt}}_{s_i} = f(U^{\text{tgt}}_{s_i}) \quad (2)$$

where f is a personality predictor.

For each trait $d \in D = \{O, C, E, A, N\}$, we aggregate speaker-level scores as:

$$\mathbf{P}^{\text{src}}_d = \{p^{\text{src}}_{s_i,d}\}_{s_i \in S} \quad (3)$$

$$\mathbf{P}^{\text{tgt}}_d = \{p^{\text{tgt}}_{s_i,d}\}_{s_i \in S} \quad (4)$$

where $p^{\text{src}}_{s_i,d}$ and $p^{\text{tgt}}_{s_i,d}$ denote the d-th trait score of speaker s_i in source and target text, respectively.

The personality consistency of translation system T is then measured by comparing $\{\mathbf{P}^{\text{src}}_d\}_{d \in D}$ and $\{\mathbf{P}^{\text{tgt}}_d\}_{d \in D}$.

Key Constructs. We measure *perceived, LLM-inferred character-personality consistency*, not psychometric personality; for brevity, we refer to it as *personality consistency* in the rest of the paper. Table 1 distinguishes the constructs this paper uses.

Construct	Role in this paper
Personality (psychometric)	Personality as measured by validated psychometric instruments. Not in scope: we do not claim to measure real people's personality.
Perceived character personality	The Big Five (John and Srivastava, 1999) impression an observer or model forms of a fictional character from dialogue. This is what we measure; PDB and Fandom supply noisy fan-derived labels for it.
Character voice (speaker voice)	A character's recognizable manner of speaking.
Style, register, politeness, stance	Distinct linguistic features (Biber and Conrad, 2009; Brown and Levinson, 1987; Du Bois, 2007) that may shape a character's perceived personality.
Character identity and names	Information linking dialogue to a character; may also trigger pretrained knowledge.
Named, Shuffled, Anonymized, name-only	Evaluation conditions: dialogue with the original character name; dialogue with a wrong but consistent name; dialogue with the explicit name masked (our text-only setting); name alone with no dialogue.

Table 1: Constructs used in this paper. We separate perceived character personality from linguistic features and from psychometric personality.

3.2 Consistency Metric

We measure personality consistency using **Spearman rank correlation** (ρ), which captures the monotonic relationship between source and target personality scores across speakers without assuming normality or linearity, computed per trait then averaged:

$$\bar{\rho} = \frac{1}{5} \sum_{d \in D} \rho(\mathbf{P}^{\text{src}}_d, \mathbf{P}^{\text{tgt}}_d) \quad (5)$$

We report Spearman as the primary metric throughout. Pearson correlation (r) and Mean Absolute Error (MAE) are defined and reported as supplementary metrics in Appendix I.

3.3 Personality Predictor

We instantiate the personality predictor f (§3) as an LLM-based predictor ("decoder predictor"). Concretely, given a speaker's utterance set U_{s_i}, we randomly sample a subset $\tilde{U}_{s_i} \subseteq U_{s_i}$ of up to N utterances and prompt the LLM to produce a structured personality prediction:

$$\mathbf{p}_{s_i} = f(\tilde{U}_{s_i}), \quad \tilde{U}_{s_i} \subseteq U_{s_i},\ |\tilde{U}_{s_i}| \leq N \quad (6)$$

where $\mathbf{p}_{s_i} \in [0,1]^5$ is the predicted Big Five vector.

We chose a decoder-based predictor over fine-tuned encoder models (e.g., XLM-RoBERTa) for three reasons: (1) fine-tuning an encoder predictor requires per-language personality-annotated dialogue, which is not available for most of our target languages; (2) a multilingual LLM can evaluate across multiple languages without per-language



Figure 1: Overview of our cross-lingual personality consistency evaluation framework. Source-language dialogue (English GAME OF THRONES, Japanese ONE PIECE) is translated into target languages by various MT systems. An LLM-based personality predictor estimates Big Five traits for each speaker in both source and translated text, and personality consistency is measured by comparing these traits via Spearman rank correlation.

fine-tuning; (3) the LLM's structured output enables consistent scoring across experiments. The full evaluation prompt is provided in Appendix B.

4 Experimental Settings

4.1 Datasets

We use two datasets spanning two source languages.

Game of Thrones (En). Our main multilingual setting is an English subtitle corpus from *Game of Thrones*, with 1,100 utterances across 11 evaluated characters. Both the English source subtitles and the 12 target-language subtitles (used as the human translation reference) are downloaded from OpenSubtitles,[1] with speaker labels obtained from the Bazinga speaker-tagged dialogue dataset (Lerner et al., 2022). We normalize speaker names, remove non-dialogue markup, and retain main characters with sufficient dialogue for stable speaker-level prediction (full pipeline in Appendix C).

One Piece (Ja). We construct a Japanese–English parallel corpus from the anime *One Piece*, containing 6,465 aligned utterances across 15 characters. Both Japanese and English subtitles are sourced from OpenSubtitles; speaker labels (character names) are obtained from the transcript dataset of Turan (2025). Because the original subtitles lack speaker annotations, we use these externally sourced speaker tags as the alignment anchor. Construction proceeds in three stages (Appendix C): (1) timestamp-based alignment of Japanese subtitle files with English transcripts, (2) LLM-based quality filtering to classify each pair as MATCH, PARTIAL, or MISMATCH, and (3) LLM-based correction of mismatched pairs using contextual re-selection. This dataset enables evaluation of the Ja→En direction.

Ground Truth Personality Scores. We obtain ground-truth (GT) Big Five scores from the Personality Database (PDB),[2] a crowdsourced platform where users vote on the personality traits of fictional characters. We collect PDB data for all main characters (11 from *Game of Thrones*, 15 from *One Piece*) using an automated web scraper that extracts vote distributions for each Big Five dimension (Appendix C).[3] To obtain continuous trait scores, we treat each vote distribution as an empirical histogram and compute its weighted average (expectation), then normalize to $[0, 1]$. To investi-

[1] https://www.opensubtitles.org/

[2] https://www.personality-database.com

[3] PDB vote distributions used in this paper are snapshots retrieved on 2026-05-17 (*Game of Thrones*) and 2025-12-18 (*One Piece*). Because PDB scores reflect ongoing community voting, the values may change over time.

gate whether PDB reflects a stable personality signal and not a single-source artifact, we additionally build an independent Big Five ground-truth from each character's Fandom-wiki biography, scored by a large language model with no access to PDB or the dialogue corpora (construction details in Appendix C); PDB–wiki character-level agreement is reported in §6.1.

4.2 Translations

We evaluate a human reference condition and five MT systems spanning two categories:

1. **Human Translation (Reference)**: We use existing subtitle translations from OpenSubtitles, aligned line by line with the source subtitles. For *Game of Thrones*, these are the subtitles in the 12 target languages; for *One Piece*, they are the fan-made English subtitles that also form the English side of the parallel corpus (§4.1). Human translation is evaluated with the same personality-consistency pipeline.

2. **Dedicated MT APIs**: Google Translate and DeepL, which take raw text as input without prompt control.

3. **LLMs**: GPT-4o (via OpenAI API), Qwen3-32B (Yang et al., 2025), and Aya-expanse-32B (Dang et al., 2024). All three use the same translation prompt template (Appendix B.2).

4.3 Language Pairs

We evaluate on translations of *Game of Thrones* from English into 12 target languages (Arabic, German, Spanish, French, Hindi, Japanese, Korean, Portuguese, Russian, Turkish, Vietnamese, Chinese) and additionally evaluate the Japanese→English direction on *One Piece*. All language pairs are reported in the main results (Table 2). Note that DeepL does not support Hindi or Vietnamese and thus covers 10 of the 12 languages.

4.4 Personality Predictor

We use Qwen3-32B (Yang et al., 2025) as the primary predictor for all main experiments. To assess predictor robustness, we also use GPT-4o (`gpt-4o-2024-08-06`) as a secondary predictor via the OpenAI API for cross-predictor agreement analysis (§6.3). Both predictors are run with temperature 0. For each speaker, we concatenate up to $N = 100$ randomly sampled utterances into a single input, and the predictor returns one Big Five vector per speaker; this value is selected based on the context-length sensitivity analysis (§6.4). Unless otherwise noted, we use a fixed Chinese-language evaluation prompt with three few-shot examples drawn from speakers outside our evaluation sets (details in Appendix B.1), shared across all conditions. We chose Chinese over English for the evaluation prompt because a non-English instruction language reduces the influence of the predictor's English-centric pretraining priors and encourages reliance on the input text; a prompt-language diagnostic supporting this choice is reported in Appendix L, and a few-shot vs. zero-shot comparison in Appendix K. The prompt includes the speaker's name as an identifier (e.g., "Tyrion," "Luffy") in the input header, followed by their numbered utterances.

5 Results

5.1 Personality Consistency

All results in this section use the original character names in the evaluation prompt (the **Named** condition). Table 2 presents personality consistency across the 12 *Game of Thrones* language pairs and the *One Piece* Ja→En direction for human translation together with five MT systems.

Two points are visible in the consistency results. First, all systems achieve moderate-to-strong measured personality consistency (MT systems span $\bar{\rho} = 0.70$–0.96 across cells; human translation spans 0.54–0.91; per-trait breakdowns in Appendix J) under the Named evaluation condition.

Second, **no single system consistently dominates**: the numerically highest condition varies by direction, including both human and MT conditions. Translation shifts the measured trait scores away from their source values, and the size of this shift depends on the translation system; we illustrate this for *One Piece* in Appendix D, with supporting data in Appendices G and H. The numerically best system varies with the target language: of the 12 *Game of Thrones* directions, Qwen3 is highest in four, Google, GPT-4o, and Aya-expanse in two each, and Human and DeepL in one each. The mean-row bootstrap intervals (95% CI, resampling the 12 target languages, Table 2) overlap substantially across systems, reinforcing that no single system separates from the rest. Utterance sampling adds further uncertainty: re-drawing the utterances of each character 20 times on *One Piece* Ja→En

Direction	Dataset	Human	Google	DeepL	GPT-4o	Qwen3	Aya-expanse
En→Ar	GoT (11 chars)	0.826	0.906	0.802	**0.912**	0.894	0.893
En→De		0.819	**0.870**	0.853	0.797	0.755	0.867
En→Es		0.846	0.782	**0.889**	0.880	0.792	0.807
En→Fr		0.905	0.787	0.877	**0.916**	0.879	0.854
En→Hi		0.543	0.787	N/A	0.777	**0.822**	0.695
En→Ja		0.812	0.854	0.860	0.822	**0.893**	0.805
En→Ko		0.848	0.832	0.864	0.861	0.858	**0.909**
En→Pt		0.838	0.909	0.830	0.771	0.916	**0.917**
En→Ru		0.751	0.800	0.790	0.846	**0.863**	0.737
En→Tr		0.796	0.818	0.856	0.817	**0.917**	0.887
En→Vi		0.826	**0.956**	N/A	0.914	0.921	0.945
En→Zh		**0.883**	0.801	0.789	0.840	0.822	0.827
GoT mean [95% CI]		0.808 [.75,.85]	0.842 [.81,.87]	0.841 [.82,.86]	0.846 [.82,.87]	**0.861** [.83,.89]	0.845 [.80,.88]
Ja→En	One Piece (15 chars)	0.803	0.757	0.790	**0.873**	0.738	0.830
One Piece Ja→En [95% CI]		0.803 [.64,.91]	0.757 [.51,.88]	0.790 [.62,.88]	**0.873** [.75,.92]	0.738 [.50,.85]	0.830 [.68,.90]

Table 2: Personality consistency (mean Spearman $\bar{\rho}$ across five Big Five traits) for human translation together with five MT systems. Predictor: Qwen3-32B, few-shot. GoT = *Game of Thrones*; Human = the existing target-language OpenSubtitles subtitle translation. DeepL covers 10 of the 12 languages because Hindi and Vietnamese are unsupported (N/A). The GoT mean row's 95% interval is obtained by resampling the 12 target languages with replacement (10,000 times) and recomputing the mean, computed separately per system. The One Piece Ja→En interval instead resamples the 15 characters with replacement (10,000 times, percentile bootstrap), since this direction has no other language to resample over. Bold indicates the numerically highest condition per direction.

moves individual systems by up to 0.06 and narrows the gaps between them, and thus the system rankings in Table 2 are uncertain (Appendix I, Table 14). Open-source LLMs show particularly high variance (e.g., Aya-expanse ranges from 0.695 to 0.945 across *Game of Thrones* pairs).

To estimate how much of this measured consistency comes from the dialogue itself rather than from the character's name, Table 3 compares Named, Shuffled, and Anonymized conditions. Anonymizing the names lowers measured consistency on both corpora, uniformly on *Game of Thrones* (-42% to -52% across the six systems) and system-dependently on *One Piece* (-8% to -51%). On *Game of Thrones*, Shuffled names (wrong but consistent across source and target) stay within 0.09 of Named, whereas anonymization roughly halves consistency, and thus the name mainly acts as a stable identity anchor shared by the source- and target-side predictions. Part of the measured consistency therefore comes from name-based priors, and the size of this part depends on how familiar the predictor is with the cast. The name-only condition measures this directly (Appendix A): from the name alone, Qwen3-32B reaches a ground-truth agreement of $\rho = 0.50$ on *Game of Thrones* and 0.27 on *One Piece*, and GPT-4o, which knows both casts, reaches 0.77 and 0.71. Anonymization masks only the explicit name, and distinctive lines can still identify a character (see Limitations).

System	Named	Shuffled	Anon.	$\Delta\%$
Game of Thrones (12-lang mean, 11 characters)				
Human	0.808	0.780	0.403	-50%
Google	0.842	0.788	0.428	-49%
DeepL	0.841	0.776	0.400	-52%
GPT-4o	0.846	0.793	0.489	-42%
Qwen3	0.861	0.773	0.442	-49%
Aya-expanse	0.845	0.793	0.453	-46%
One Piece (Ja→En, 15 characters)				
Human	0.803	0.667	0.409	-49%
Google	0.757	0.629	0.371	-51%
DeepL	0.790	0.790	0.645	-18%
GPT-4o	0.873	0.682	0.806	-8%
Qwen3	0.738	0.579	0.383	-48%
Aya-expanse	0.830	0.791	0.614	-26%

Table 3: Effect of name manipulation on personality consistency (mean Spearman $\bar{\rho}$; Qwen3-32B predictor, few-shot; conditions defined in Table 1). *Game of Thrones*: mean over all 12 language pairs (DeepL covers 10), 11 characters; *One Piece*: Ja→En, 15 characters. $\Delta\%$ is the change from Named to Anonymized.

5.2 Reference-Based Quality Metrics Do Not Capture Personality Consistency

Does a translation that better matches a human reference also score higher on personality consistency? In our setup, not reliably. Table 4 compares translation quality (BLEU, COMET) with personality consistency across all MT systems.

On *Game of Thrones*, Google attains the highest mean BLEU (16.26) and DeepL the highest mean COMET (0.722), while Qwen3 attains the

System	BLEU	COMET	$\bar{\rho}$
Game of Thrones En→12 languages (mean)			
Google	**16.26**	0.713	0.842
DeepL	15.09	**0.722**	0.841
GPT-4o	16.15	0.715	0.846
Qwen3	13.86	0.702	**0.861**
Aya-expanse	14.89	0.708	0.845
One Piece Ja→En (ref = Human En, N=6,465)			
Google	9.84	0.671	0.757
DeepL	12.33	0.683	0.790
GPT-4o	**12.99**	**0.686**	**0.873**
Qwen3	11.64	0.675	0.738
Aya-expanse	11.93	0.676	0.830

Table 4: Translation quality (BLEU, COMET) and personality consistency ($\bar{\rho}$, from Table 2) for five MT systems. Bold indicates best per column per direction.

highest personality consistency (0.861) despite a lower BLEU (13.86); on *One Piece*, GPT-4o leads both quality and consistency. Within each language pair, neither BLEU nor COMET tracks personality consistency across the available MT systems: the mean within-language Spearman is near zero (BLEU $\bar{\rho} = -0.08$; COMET $\bar{\rho} = -0.19$), and a one-sample t-test over the 13 within-language ρ values does not reject zero (BLEU p=0.59, COMET p=0.21). BLEU and COMET measure similarity to human references, yet human translators themselves perform cultural adaptation that may reduce personality consistency; the weak association may thus be partly an artifact of reference-based evaluation.

This adaptation is quantifiable: in an auxiliary En→Ja subtitle analysis (Appendix F), human subtitles compress and rewrite personality-relevant surface pragmatics more aggressively than MT, and either removing source-visible cues or injecting target-specific characterization lowers source–target consistency without making the translation worse overall. Human translation is thus a reference condition rather than an upper bound under our metric; whether its deliberate adaptation is desirable depends on the intent of the localization. The metric makes the size of the shift visible and flags shifts that are not deliberate: the MT case study in Appendix D traces the shift to softened register cues. BLEU and COMET detect neither kind of shift, and thus we recommend reporting personality consistency alongside reference-based quality metrics when evaluating character-driven localization.

Corpus	Chars	r	95% CI
Game of Thrones	11	0.78	[0.60, 0.88]
One Piece	15	0.82	[0.75, 0.89]

Table 5: Character-level Big Five agreement between PDB (crowd votes) and an independent source obtained by scoring each character's Fandom-wiki *Personality* section with a large language model (GPT-5.4, three runs averaged), with no access to PDB or the dialogue corpora. Pearson r over all (character, trait) pairs; 95% CI from resampling characters with replacement (10,000 times).

6 Analysis and Discussion

6.1 Ground-Truth Validity

Our personality ground-truth is the Personality Database (PDB), a crowd-sourced platform. The prediction accuracy reported next (§6.2) measures agreement with these labels, and thus it is informative only if the labels themselves are reliable rather than an idiosyncrasy of a single source. We therefore first validate the labels: we compare PDB against a second, independently constructed ground-truth: Big Five scores assigned by a large language model to each character's Fandom-wiki biography's *Personality* section, with no access to PDB or to the dialogue corpora (construction details in Appendix C). For our two corpora we correlate the two five-dimensional profiles over all evaluated characters (*Game of Thrones* and *One Piece*, both fully covered by the two sources). Table 5 reports the Pearson correlation. Because each corpus has only a handful of characters, we attach a 95% confidence interval estimated by bootstrapping. We redraw the character set with replacement 10,000 times, keeping each character's five trait values together in each redraw, recompute the correlation on every redrawn set, and take the 2.5th and 97.5th percentiles of the resulting 10,000 values as the interval. The two ground-truths show strong character-level agreement on both main corpora (Pearson $r = 0.78$ for *Game of Thrones* and $r = 0.82$ for *One Piece*), with both confidence intervals well above zero. PDB and an independently constructed Big Five source thus capture a shared personality signal, and the evaluation does not rest on a single ground-truth. Per-character ground-truth values are listed in Appendix M. Both sources are fandom-derived: their agreement shows convergence between two independent representations of audience-perceived character personality, not psy-

	src.	tgt mean					
Dir.	**Ori.**	**Human**	**Google**	**DeepL**	**GPT-4o**	**Qwen3**	**Aya-expanse**
Game of Thrones (11 characters, 12 languages)							
GoT mean	0.603	0.528	0.569	**0.577**	0.558	0.556	0.562
One Piece (15 characters)							
Ja→En	0.596	**0.661**	0.652	0.640	0.618	0.629	0.595

Table 6: Personality prediction accuracy vs. PDB ground truth (ρ) for source and target languages. Predictor: Qwen3-32B, few-shot. Bold indicates best target per direction. All six systems share one *Game of Thrones* source-side score because the English source text is identical.

chometric validity. For our construct this is the relevant target: entertainment translation cares about the personality audiences perceive, PDB and Fandom aggregate exactly such audience judgments, and since a fictional character cannot complete a self-report questionnaire, observer reports are the natural source, an established one in personality measurement (Ashton and Lee, 2025). Appendix N adds an external, fandom-independent check of the predictor itself.

6.2 Personality Prediction Accuracy

Table 6 reports Spearman correlation between predicted Big Five scores and PDB ground truth, separately for source-language (src) and target-language (tgt) predictions. For Ja→En, target predictions improve over source for five of six systems (e.g., Human $0.596 \rightarrow 0.661$); Aya-expanse is essentially unchanged ($0.596 \rightarrow 0.595$). For *Game of Thrones*, all six systems share the same source-side score (0.603); target-side accuracy degrades for every system, but by different amounts, with DeepL retaining the most (a drop of 0.026) and Human the least (a drop of 0.075). Prediction-accuracy rankings differ from consistency rankings: on *Game of Thrones*, DeepL tops target accuracy (0.577) whereas Qwen3 tops consistency (0.861); on *One Piece*, Human tops target accuracy (0.66) but GPT-4o tops consistency (0.87). The two metrics need not move together because they are correlations against different references: consistency measures agreement between target- and source-language predictions, while accuracy measures agreement between the target-language prediction and an external, only moderately source-aligned ground truth (One Piece's Japanese-source accuracy against PDB is itself just 0.60, not 1.0). A translation's deviation from the source prediction can therefore move the target-language prediction closer to the ground truth than the source was (raising accuracy while lowering consistency, as for Human on One Piece) or further from it (lowering both, as for Human on *Game of Thrones*, whose target accuracy of 0.528 and consistency of 0.808 are both the lowest of the six systems). Since PDB annotations are predominantly contributed by English-speaking users, these accuracy scores reflect agreement with English-centric perception rather than language-neutral ground truth.

System	**GoT (12 langs)**	**One Piece**
Human	0.862	0.918
Google	0.866	0.796
DeepL	**0.892**	0.913
GPT-4o	0.873	0.909
Qwen3	0.867	0.806
Aya-expanse	0.869	**0.932**

Table 7: Personality consistency ($\bar{\rho}$, few-shot) using GPT-4o as predictor (cf. Table 2 for Qwen3-32B).

6.3 Cross-Predictor Agreement

We re-evaluate all systems using GPT-4o as a second predictor (Table 7). Both predictors agree on the core finding: system rankings are configuration-sensitive, and no single system consistently dominates.

On *Game of Thrones*, GPT-4o yields uniformly higher and more compressed correlations ($\bar{\rho} = 0.86$–0.89) than Qwen3 (0.81–0.86). This pattern alone does not identify a cause: it could reflect calibration, output granularity, or cross-lingual stability differences between the evaluators; the name-only diagnostic (Appendix A) provides the separate, direct evidence on pretrained character knowledge. The numerically best system differs by predictor (Qwen3 under the Qwen3 predictor, DeepL under GPT-4o), indicating that specific system rankings are not predictor-invariant. Few-shot vs. zero-shot prompting can likewise alter specific rankings (Appendix K). While the numerically best system varies across predictor, prompt, and language pair, the conclusion that no single system consistently

System	GoT (12 langs)	One Piece
Human	0.710	0.739
Google	0.695	0.709
DeepL	0.688	0.728
GPT-4o	0.693	**0.743**
Qwen3	**0.711**	0.699
Aya-expanse	0.703	0.741

Table 8: Target-language prediction accuracy vs. PDB ground truth (ρ) using GPT-4o as predictor (cf. Table 6 for Qwen3-32B). Bold indicates best per corpus. On both corpora the numerically best system depends on the predictor.

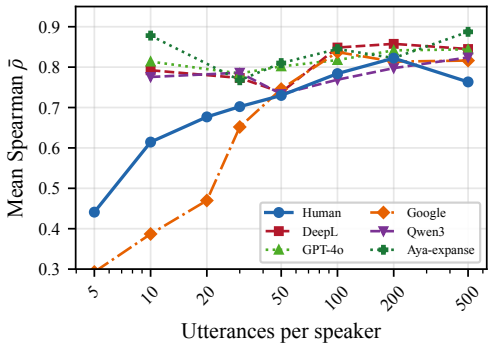


Figure 2: Effect of utterance count on personality consistency (Ja→En, 15 characters). Predictor: Qwen3-32B, few-shot.

dominates holds across the tested configurations.

The same predictor swap affects target-language prediction accuracy vs. PDB (Table 8): under GPT-4o the absolute scores rise and compress, and the numerically top systems differ from those under Qwen3 (Table 6). On *Game of Thrones* the top system changes with the predictor (DeepL under Qwen3-32B, Qwen3 under GPT-4o, the latter narrowly over Human at 0.710); on *One Piece*, Human leads under Qwen3, whereas GPT-4o (0.743) narrowly leads Aya-expanse (0.741) and Human (0.739) under the GPT-4o predictor. Accuracy rankings are therefore predictor-dependent diagnostics rather than definitive system rankings.

Appendices D and E report case studies and dialogue-context translation.

6.4 Context-Length Sensitivity

How many utterances per speaker are needed for stable prediction? Figure 2 shows the effect of varying utterance count from 5 to 500 on personality consistency for Ja→En (One Piece, 15 characters).

Consistency rises with utterance count and plateaus around 100 across systems. The rise is steepest for human translation (from $\bar{\rho} = 0.44$ at 5 utterances to 0.82 at 200) and for Google; the remaining systems (DeepL, GPT-4o, Qwen3, Aya-expanse) are already near their plateau by 30–50 utterances. The marginal gain beyond 100 utterances is negligible, and many speakers in our datasets have fewer than 200 utterances available; additional lines tend to dilute rather than reinforce speaker-specific cues. We therefore use $N = 100$ **utterances per speaker** as a practical threshold that captures most of the signal while remaining feasible across all language pairs.

7 Conclusion

We presented a framework for quantifying perceived, LLM-inferred character-personality consistency in translation, evaluated on two main character-rich corpora, *Game of Thrones* (English→12 target languages, 11 characters) and *One Piece* (Japanese→English, 15 characters), via speaker-level personality prediction and trait-wise consistency aggregation. Two findings stand out: (1) personality consistency is only weakly associated with BLEU/COMET, indicating that perceived-personality consistency is a distinct evaluation dimension; (2) no translation system is a reliable top performer across languages, corpora, and predictor/prompt configurations. We release our evaluation framework to facilitate future research.[4]

Two promising directions are *personality-aware translation* conditioned on speaker Big Five profiles, and cross-cultural analysis of trait perception.

[4]Code and evaluation tools will be released at https://github.com/ku-nlp/character-personality-mt.

Limitations

Predictor Coverage. Our conclusions rely on two LLM predictors (Qwen3-32B and GPT-4o). Both are large-scale models with extensive pretraining on multilingual data; whether smaller or domain-specific models exhibit the same patterns (e.g., name dependence, cross-lingual prior asymmetry) remains untested.

Ground-Truth Labels. PDB and Fandom-wiki are fandom-derived labels of audience-perceived character personality, not psychometrically validated measurements. They may be shaped by English-centric cultural framing and character popularity, and prediction accuracy against them means agreement with audience perception, not with "true" personality.

Scope and Generality. Our translation evidence is limited to two well-known fictional works with speaker labels, with small speaker-level samples (11 *Game of Thrones* and 15 *One Piece* characters), and we do not claim that the results generalize to unknown speakers, real-world dialogue, less-famous characters, or other domains. The external predictor check (Appendix N) uses non-famous actors but is acted, untranslated dialogue and cannot establish this generalization; more translated datasets, especially with less-famous characters, are needed. The framework itself also has scope requirements: it needs stable speaker groups, enough dialogue per speaker, and characters whose personalities are intentionally developed; it is not designed for genres where character personality is weak or irrelevant, and we do not claim that it applies to all translation tasks. The *One Piece* alignment pipeline additionally uses LLM-based filtering and correction; our manual check of 100 stratified pairs found 99 correctly aligned (Appendix C).

Name Priors and Anonymization. Named results draw on pretrained knowledge linked to famous character names, and anonymization masks explicit names only: distinctive catchphrases and recognizable lines remain in the dialogue, so the Anonymized condition is not a perfect text-only control. System rankings also change when the predictor changes (§6.3), and we present rankings as results under the stated naming and evaluator conditions.

No Direct Human Validation. We have not asked human judges to compare source and translated dialogue and rate whether character voice is preserved; the source-to-target consistency score itself therefore has no direct human validation. Appendix N validates the predictor against human personality annotations at the monolingual level only. Direct human validation of the cross-lingual score is important future work.

Ethical Considerations

Data. We use publicly available or subtitle-derived dialogue data for GAME OF THRONES and ONE PIECE. Big Five personality ground truth is scraped from the Personality Database, a public crowdsourced platform. No personally identifiable information is collected or processed; all speakers analyzed are fictional. We will release derived evaluation metadata and personality annotations where redistribution is permitted, and otherwise release scripts and alignment metadata to support reproducibility.

Potential Misuse. Our framework measures personality *consistency* across languages, not personality *detection* of real individuals. We caution against applying this methodology to infer personality traits of real people from their text, as LLM-based personality prediction lacks psychometric validity and could be used to discriminate against or stereotype individuals.

Copyright. The subtitle-derived corpora contain copyrighted dialogue from entertainment media. We use short, character-attributed excerpts for research evaluation, do not release copyrighted subtitle text when redistribution is not permitted, and release only derived metadata, code, and pointers where appropriate.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP23K28144 and JST BOOST Grant Numbers JPMJBS2407, JPMJBY25D2. We thank the anonymous reviewers for their constructive comments, which helped improve this work. We also appreciate the careful attention of the meta reviewer.

References

Jacopo Amidei, Jose Gregorio Ferreira De Sá, Rubén Nieto Luna, and Andreas Kaltenbrunner. 2025. Exploring the impact of language switching on personality

traits in LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2370–2378.

Shlomo Argamon, Moshe Koppel, James W Pennebaker, and Jonathan Schler. 2009. Automatically profiling the author of an anonymous text. *Communications of the ACM*, 52(2):119–123.

Michael C. Ashton and Kibeom Lee. 2025. Self- and observer reports of personality. *Annual Review of Psychology*, 76:771–795.

Douglas Biber and Susan Conrad. 2009. *Register, Genre, and Style*. Cambridge University Press.

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

Sirou Chen, Sakiko Yahata, Shuichiro Shimizu, Zhengdong Yang, Yihang Li, Chenhui Chu, and Sadao Kurohashi. 2024. MELD-ST: An emotion-aware speech translation dataset. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 10118–10126, Bangkok, Thailand. Association for Computational Linguistics.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya Expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

John W. Du Bois. 2007. The stance triangle. In Robert Englebretson, editor, *Stancetaking in Discourse: Subjectivity, Evaluation, Interaction*, pages 139–182. John Benjamins.

Yahui Fu, Chenhui Chu, and Tatsuya Kawahara. 2024. StyEmp: Stylizing empathetic response generation via multi-grained prefix encoder and personality reinforcement. In *Proceedings of the 25th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 172–185.

Yuan Gao, Hao Shi, Yahui Fu, Chenhui Chu, and Tatsuya Kawahara. 2025. Bridging speech emotion recognition and personality: Dataset and temporal interaction condition network. *Preprint*, arXiv:2505.13978.

Lewis R. Goldberg. 1990. An alternative "description of personality": The Big-Five factor structure. *Journal of Personality and Social Psychology*, 59(6):1216–1229.

Guangyuan Jiang, Manjie Xu, Song-Chun Zhu, Wenjuan Han, Chi Zhang, and Yixin Zhu. 2023. Evaluating and inducing personality in pre-trained language models. *Advances in Neural Information Processing Systems*, 36:10622–10643.

Hang Jiang, Xianzhe Zhang, and Jinho D Choi. 2020. Automatic text-based personality recognition on monologues and multiparty dialogues using attentive networks and contextual embeddings (student abstract). In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13821–13822.

Oliver P. John and Sanjay Srivastava. 1999. The Big Five trait taxonomy: History, measurement, and theoretical perspectives. In Lawrence A. Pervin and Oliver P. John, editors, *Handbook of Personality: Theory and Research*, 2nd edition, pages 102–138. Guilford Press, New York.

Seungbeen Lee, Seungwon Lim, Seungju Han, Giyeong Oh, Hyungjoo Chae, Jiwan Chung, Minju Kim, Beong-woo Kwak, Yeonsoo Lee, Dongha Lee, Jinyoung Yeo, and Youngjae Yu. 2025. Do LLMs have distinct and consistent personality? TRAIT: Personality testset designed for LLMs with psychometrics. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 8412–8452.

Paul Lerner, Juliette Bergoënd, Camille Guinaudeau, Hervé Bredin, Benjamin Maurice, Sharleyne Lefevre, Martin Bouteiller, Aman Berhe, Léo Galmant, Ruiqing Yin, and Claude Barras. 2022. Bazinga! A dataset for multi-party dialogues structuring. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3434–3441, Marseille, France. European Language Resources Association.

Jiwei Li, Michel Galley, Chris Brockett, Georgios Spithourakis, Jianfeng Gao, and Bill Dolan. 2016. A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 994–1003.

Wenkai Li, Jiarui Liu, Andy Liu, Xuhui Zhou, Mona T. Diab, and Maarten Sap. 2025. BIG5-CHAT: Shaping LLM personalities through training on human-grounded data. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 20434–20471.

François Mairesse, Marilyn A Walker, Matthias R Mehl, and Roger K Moore. 2007. Using linguistic cues for the automatic recognition of personality in conversation and text. *Journal of Artificial Intelligence Research*, 30:457–500.

Robert R. McCrae and Paul T. Costa, Jr. 1997. Personality trait structure as a human universal. *American Psychologist*, 52(5):509–516.

Paul Michel and Graham Neubig. 2018. Extreme adaptation for personalized neural machine translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 312–318.

Shachar Mirkin, Scott Nowson, Caroline Brun, and Julien Perez. 2015. Motivating personality-aware

machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1102–1108.

Jeremy Munday, Sara Ramos Pinto, and Jacob Blakesley. 2022. *Introducing Translation Studies: Theories and Applications*, 5th edition. Routledge, London.

Eugene Albert Nida. 1964. *Toward a Science of Translating: With Special Reference to Principles and Procedures Involved in Bible Translating*. E. J. Brill, Leiden.

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.

Ella Rabinovich, Raj Nath Patel, Shachar Mirkin, Lucia Specia, and Shuly Wintner. 2017. Personalized machine translation: Preserving original author traits. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1074–1084.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702.

David P Schmitt, Jüri Allik, Robert R McCrae, and Verónica Benet-Martínez. 2007. The geographic distribution of Big Five personality traits: Patterns and profiles of human self-description across 56 nations. *Journal of Cross-Cultural Psychology*, 38(2):173–212.

H. Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E. P. Seligman, and Lyle H. Ungar. 2013. Personality, gender, and age in the language of social media: The open-vocabulary approach. *PLoS ONE*, 8(9):e73791.

Gregory Serapio-García, Mustafa Safdari, Clément Crepy, Luning Sun, Stephen Fitz, Peter Romero, Marwa Abdulhai, Aleksandra Faust, and Maja Matarić. 2025. A psychometric framework for evaluating and shaping personality traits in large language models. *Nature Machine Intelligence*, 7(12):1954–1968.

Lei Sun, Jinming Zhao, and Qin Jin. 2024. Revealing personality traits: A new benchmark dataset for explainable personality recognition on dialogues. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 19988–20002.

Seungjong Sun, Seo Yeon Baek, and Jang Hyun Kim. 2025. Personality vector: Modulating personality of large language models by model merging. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 24656–24677.

Ramazan Turan. 2025. One piece transcripts with character names: Episodes 382–777. Dataset.

Longyue Wang, Chenyang Lyu, Tianbo Ji, Zhirui Zhang, Dian Yu, Shuming Shi, and Zhaopeng Tu. 2023. Document-level machine translation with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16646–16661.

Sanae Yamashita, Koji Inoue, Ao Guo, Shota Mochizuki, Tatsuya Kawahara, and Ryuichiro Higashinaka. 2023. RealPersonaChat: A realistic persona chat corpus with interlocutors' own personalities. In *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, pages 852–861.

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. *arXiv preprint arXiv:2505.09388*.

Tao Yang, Jinghao Deng, Xiaojun Quan, and Qifan Wang. 2023a. Orders are unwanted: dynamic deep graph convolutional network for personality detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13896–13904.

Tao Yang, Tianyuan Shi, Fanqi Wan, Xiaojun Quan, Qifan Wang, Bingzhe Wu, and Jiaxiang Wu. 2023b. PsyCoT: Psychological questionnaire as powerful chain-of-thought for personality detection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 3305–3320.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020*.

A Name-Condition Diagnostic

The Named/Shuffled/Anonymized comparison and its system-level result (Table 3) are presented in §5.1; the three conditions are defined in Table 1 and are applied symmetrically to both the source-side and the target-side evaluation. This appendix supplements that result with the name-only control, which isolates the name prior itself.

Name-only agreement on *Game of Thrones*. The name-only condition (Table 1; summary across corpora in Table 9) gives the predictor the character's name with every utterance withheld. It ignores the dialogue, and thus it is independent of the translation system and is run once per direction. On *Game of Thrones*, the names alone recover much of the ground-truth ranking: Qwen3-32B reaches $\rho = 0.50$ against PDB from the name alone and 0.60 with the full dialogue, and its name-only predictions agree with the full Named predictions at $\rho \approx 0.70$. Under GPT-4o the prior is stronger (0.77 name-only, 0.68 with the full dialogue, $\rho \approx 0.84$ between the name-only and full Named predictions). With 11 characters, the character-bootstrap interval for the difference between name-only and full-dialogue agreement includes zero for both predictors, and thus we report only the two levels for this corpus. The comparison shows that the name prior is substantial on *Game of Thrones* and stronger under GPT-4o than under Qwen3-32B. This motivates the choice of Qwen3-32B as the main predictor (§4.4) and explains the large Named-to-Anonymized drop in Table 3. The size of the prior depends on the evaluator's pretraining familiarity with the cast.

Name-only and text-only recoverability on One Piece. On *One Piece*, Qwen3-32B's name-only agreement with PDB ground truth is 0.27, and its rank agreement with the full Named prediction is only $\rho \approx 0.33$; adding the utterances raises ground-truth agreement to 0.59, a gain of $+0.32$ (speaker-bootstrap 95% CI $[+0.08, +0.58]$), and thus the utterances contribute substantial information beyond the name on this corpus. Under GPT-4o, by contrast, name-only already reaches 0.71 on this famous cast (Table 9), and thus we run the strictest text-only test with Qwen3-32B, whose weak name prior guarantees that even a covertly recognized character could contribute at most the name-only level: target-side ground-truth agreement under the Anonymized condition. On Ja→En, averaging the four MT systems (Google, GPT-4o, Qwen3, Aya-expanse), Anonymized source-side agreement is 0.606 and target-side agreement is 0.493 (speaker-bootstrap 95% CI $[0.29, 0.62]$, clearly above zero), and the source–target difference interval includes zero ($[-0.31, 0.09]$). Trait impressions therefore remain recoverable from anonymized translated text alone, although the 15-character sample gives wide intervals.

Corpus	Predictor	name-only	+ utterances
Game of Thrones (11)	Qwen3-32B	.50	.60
Game of Thrones (11)	GPT-4o	.77	.68
One Piece (15)	Qwen3-32B	.27	.59
One Piece (15)	GPT-4o	.71	.71

Table 9: Name-only agreement with PDB ground truth versus agreement when the utterances are included, per corpus (roster size) and evaluator; source-side Named full text throughout. Name-only agreement is low where the evaluator lacks pretrained knowledge of the cast (Qwen3-32B on *One Piece*), and adding the utterances raises it substantially ($+0.32$, character-bootstrap 95% CI $[+0.08, +0.58]$). In the other three rows, where the evaluator knows the cast, the names alone reach 0.50 to 0.77, name-only and full-utterance agreement differ by at most 0.10, and the character-bootstrap interval of each difference includes zero. Protocol details in the paragraphs above.

B Evaluation Prompt

The personality evaluation prompt used for all experiments (English translation of our Chinese prompt):

```
You are an expert psychologist
specializing in Big Five
personality prediction.  Based
on the following utterances by
this speaker, estimate their Big
Five trait scores (0.0-1.0, 3
decimals) and output only valid
JSON with keys:  N, E, O, A, C.
```

B.1 Few-Shot Examples

Our few-shot prompt includes three example speakers with real Personality Database (PDB) scores, chosen to span diverse Big Five profiles. These speakers do not appear in any of our evaluation sets. Each example consists of five representative utterances followed by the speaker's ground-truth Big Five vector, providing the predictor with calibration anchors before scoring the target speaker.

Example 1: Squidward Tentacles (high N, low E/A)

```
1:  Another day, another
migraine.
```

```
2: Would you two please keep
it down? I'm trying to enjoy my
day off in peace.
3: I knew I should have stayed
in bed today.
4: I hate all of you.
5: Please do not disturb me.
I'm busy practicing my clarinet.
→ {N: 0.930, E: 0.198, O: 0.415,
A: 0.168, C: 0.568}
```

Example 2: SpongeBob SquarePants (high E/A/O)

```
1: I'm ready! I'm ready! I'm
ready!
2: Good morning, world, and all
who inhabit it!
3: The best time to wear a
striped sweater... is all the
time!
4: Gary, I'm so glad you're
here! I made you your favorite
snail food!
5: You don't need a license to
drive a sandwich!
→ {N: 0.569, E: 0.963, O: 0.828,
A: 0.920, C: 0.623}
```

Example 3: Light Yagami (high C, low A/N)

```
1: This world is rotten, and
those who are making it rot
deserve to die.
2: In this world, there are
very few people who actually
trust each other.
3: You can't ever win if you're
always on the defensive. To
win, you have to attack.
4: Humans aren't made perfectly.
Everyone lies. Even so, I've
been careful not to tell lies
that hurt others.
5: Look around you, and all you
will see are people the world
would be better off without.
→ {N: 0.346, E: 0.495, O: 0.680,
A: 0.171, C: 0.946}
```

B.2 Translation Prompt for MT Systems

For LLM-based MT systems (GPT-4o, Qwen3-32B, and Aya-expanse-32B), we use the following instruction template:

```
You are a professional
translator. Translate the
following {source_lang} text to
{target_lang}. Output only the
translation, no explanations.
```

C Data Construction

PDB Personality Scraping. We collect Big Five personality ground truth for covered fictional characters by automatically scraping the Personality Database (PDB) website. Ground-truth analyses use *Game of Thrones* and *One Piece*. For each speaker profile, we extract the vote distribution across five percentage levels (0%, 25%, 50%, 75%, 100%) for each Big Five dimension. The final trait score is computed as the weighted average (expectation) over the vote histogram:

$$s_d = \frac{\sum_k p_k \cdot v_k}{\sum_k v_k} \tag{7}$$

where $p_k \in \{0, 0.25, 0.5, 0.75, 1.0\}$ is the percentage level and v_k is the number of votes at that level. Since $p_k \in [0, 1]$, the resulting scores lie in $[0, 1]$ by construction.

English-Source Subtitle Corpus. For *Game of Thrones*, English source subtitles and all 12 target-language subtitles (used as the human translation reference) are downloaded from OpenSubtitles[5] via the `subliminal` library, covering Seasons 1–6 (60 episodes per language). Speaker labels are transferred from the Bazinga speaker-tagged dialogue dataset (Lerner et al., 2022) by approximate timestamp-and-text matching. We normalize speaker names, remove non-dialogue markup and stage directions, merge duplicate or near-empty subtitle fragments, and retain characters with enough dialogue for stable speaker-level prediction. For each retained character, we cap the input at up to 100 utterances using a fixed sampling procedure, yielding 1,100 utterances across 11 characters. The English source files are translated by the MT systems and aligned with the downloaded target-language subtitles for the human-reference condition.

One Piece Corpus Construction. The Japanese–English parallel corpus is constructed in three stages:

1. **Timestamp-based alignment.** We align Japanese subtitle files (SRT format) with English transcripts using timestamp matching with a 2-second tolerance window. Speaker names are obtained from the transcript dataset of Turan (2025), which provides character-labeled transcripts for episodes 382–777, and mapped to English names. This yields an initial set of candidate Ja–En pairs.

2. **LLM-based quality filtering.** Each aligned pair is classified by an LLM (Qwen3-30B) into four categories: MATCH (same speaker,

[5] https://www.opensubtitles.org/

correct translation), PARTIAL (same speaker, merged or extra content), MISMATCH (wrong speaker or wrong translation), or UNCERTAIN (insufficient evidence). Only MATCH and PARTIAL pairs are retained directly.

3. **LLM-based correction.** For MISMATCH pairs, we provide the LLM with nearby English lines from the same episode (15 nearest by timestamp) as context, and ask it to select the correct English translation. Successfully corrected pairs are added back to the corpus.

The final corpus contains 6,465 aligned utterances across 15 characters, with speaker labels and episode metadata.

Manual Alignment Validation. Because stages 2 and 3 rely on an LLM, we manually validated a stratified sample of 100 aligned pairs, drawn with a fixed seed and stratified by character within each stratum, oversampling the processing types most exposed to alignment error: 46 LLM-corrected pairs (31% of the corpus, raised to 46% of the sample), all 14 PARTIAL-overlap pairs in the corpus, and 40 directly timestamp-matched pairs. All timestamp-matched and PARTIAL-overlap pairs are correctly aligned; 45 of the 46 LLM-corrected pairs are correctly aligned, giving 99 of 100 correct overall. The single error is a misaligned pair in the LLM-corrected stratum.

Wiki Ground-Truth Construction. For the convergent-validity check in §6.1, we construct a second Big Five ground-truth from each character's *Fandom* wiki page,[6] a fan-maintained encyclopedia. Each character page contains a long-form *Personality* section describing the character's traits in prose; for example, Luffy's entry begins "Luffy's most dominant traits are his apparent lack of common sense, his utter fearlessness and unwavering belief in himself ..." For each of the 26 covered characters (*Game of Thrones* 11, *One Piece* 15), we extract this *Personality* section and prompt GPT-5.4 with the extracted text and the same Big Five rubric used by our LLM predictor (Appendix B), instructing the model to score traits *only* from the supplied text and to ignore any prior knowledge of the character. We average three independent runs per character to reduce variance and use the resulting per-character five-trait vectors as the Wiki ground-truth. By construction this ground-truth shares neither evidence (free-form prose vs. crowd votes) nor raters (fan editors+LLM vs. PDB voters) with PDB.

[6] https://www.fandom.com/

	Src	Ja → En Translation					
	(Ja)	Human	Google	DeepL	GPT-4o	Qwen3	Aya-expanse
Hancock	.15	$.22^{+.07}$	$.30^{+.15}$	$.20^{+.05}$	.15	$.18^{+.03}$	$.20^{+.05}$
Jinbe	.80	$.85^{+.05}$	.80	$.85^{+.05}$	.80	.80	$.75^{-.05}$
Sanji	.40	$.65^{+.25}$	$\mathbf{.85}^{+.45}$	$.60^{+.20}$	$.60^{+.20}$	$\mathbf{.85}^{+.45}$	.40

Table 10: AGREEABLENESS scores (0–1) for three *One Piece* speakers, on source Japanese and each Ja→En translation. Predictor: Qwen3-32B, few-shot. Gray superscripts show the shift (Δ) from source; **bold** = $|\Delta| \geq 0.40$.

D Personality Shifts Under Translation

To show that measured consistency tracks real linguistic change, we examine AGREEABLENESS for three *One Piece* speakers (Table 10). Japanese encodes agreeableness through honorific level (*keigo*), sentence-final particles, and first-person pronoun choice (*ore*, *washi*, *warawa*), pragmatic features with no direct English equivalent. The effect is speaker- and system-dependent: Hancock's low agreeableness, carried by explicit lexical cues, survives all systems ($\Delta \leq 0.15$) and Jinbe's warmth transfers intact, whereas Sanji's rough register (*ore*, imperative forms) is softened by most systems, raising his A by up to 0.45 (Google, Qwen3) and 0.25 under human translation. Translation can thus reshape personality-relevant pragmatic features even when it is fluent and accurate. Per-speaker vectors for all *One Piece* characters are given in Appendix G, with representative Sanji utterance pairs in Appendix H.

E Dialogue-Context Translation

Translating each utterance in isolation may discard dialogue context that carries personality. We test this on *One Piece* (Ja→En, 15 characters): each MT system re-translates every utterance together with its surrounding dialogue context, and we re-measure personality consistency (Table 11). Context yields a small, uniform gain: all five systems improve (mean $\Delta\bar{\rho} = +0.037$), and the system ranking is essentially unchanged (Spearman $\rho = 0.90$ between the utterance-level and context-level system orderings). Dialogue context thus adds a minor additive signal but does not alter system comparisons, and the utterance-level protocol is adequate for our analysis.

System	Utterance	+ Context	Δ
Google	0.757	0.790	+0.033
DeepL	0.790	0.875	+0.085
GPT-4o	0.873	0.875	+0.002
Qwen3	0.738	0.785	+0.047
Aya-expanse	0.830	0.850	+0.020
Mean	0.798	0.835	+0.037

Table 11: Personality consistency ($\bar{\rho}$) under utterance-level vs. dialogue-context translation, *One Piece* Ja→En (15 characters, Qwen3-32B predictor, few-shot). Δ is the context-minus-utterance change per system; all five gains are positive (mean +0.037).

F Human-Subtitle Rewriting of Pragmatic Cues

In an auxiliary analysis of the MELD-ST (Chen et al., 2024) En→Ja dialogue corpus (six main speakers), human subtitles average 11.3 Japanese characters per utterance against 23.3 for Google MT, contain far fewer exclamation marks (3.7 vs. 38.9 per 100 utterances), retain the English discourse marker *oh* in only 14% of utterances against 78% for MT, and assign persona-bearing Japanese first-person pronouns (*ore*, *boku*) to different characters where MT uses *watashi* uniformly. Human subtitle translation thus both removes source-visible cues and injects target-language-specific characterization, and either operation lowers source–target consistency (§5.1) without making the translation worse overall.

G One Piece Speaker-Level Vectors

Table 12 lists the per-speaker Big Five vectors underlying the *One Piece* personality-shift analysis in Appendix D (Ja source vs. Human En translation).

H Sanji Case Study: Explaining the Largest System-Dependent Agreeableness Shift

This appendix focuses on Sanji, the speaker in Table 10 with the largest system-dependent AGREEABLENESS shift. We use Google as one illustrated high-shift system ($\Delta A = +0.45$) and DeepL and GPT-4o as lower-shift contrasts (+0.20), and ask what is visible at the utterance level that is consistent with this aggregate pattern.

Table 13 lists three representative Sanji utterances together with the human English subtitle (the reference condition) and the DeepL, GPT-4o, and Google MT outputs.

Sanji's source register is signalled in Japanese by the rough first-person pronoun *ore*, abrupt imperative forms, and casual contractions such as *urusee* for *urusai* and *hanaren ja nee zo* for *hanareru na yo*. Since English normally neutralizes *ore* as "I," the observable cross-system differences are mainly in imperative force and casualness. DeepL and GPT-4o both preserve rough command-like force: *urusee* is rendered as "Shut up!" under both, and the abrupt order *hanaren ja nee zo* is rendered with the strong "don't you dare" construction under both ("don't you dare leave my side" vs. "don't you dare leave me"). The human reference also preserves this force with the idiomatic command "Can it!" in the first row, but is milder in the final row ("Don't fall behind"); this mixed pattern is consistent with its intermediate Sanji shift in Table 10. Google, in contrast, flattens *urusee* to "Noisy!" (a descriptive adjective rather than a command) and *hanaren ja nee zo* to "don't leave me" (no imperative-strengthener). The visible contrasts are consistent with the smaller Sanji shifts for DeepL/GPT-4o and the larger shift for Google in Table 10.

I Supplementary Metrics and Additional Results

In addition to Spearman rank correlation (§3), we report two supplementary metrics:

Pearson Correlation. Pearson correlation (r) measures the linear relationship between source and target personality scores, averaged across traits:

$$\bar{r} = \frac{1}{5} \sum_{d \in D} r(\mathbf{P}_d^{\text{src}}, \mathbf{P}_d^{\text{tgt}}) \tag{8}$$

Unlike Spearman, Pearson assumes a linear relationship and is sensitive to outliers and non-normality.

Mean Absolute Error. MAE quantifies the average absolute deviation between source and target trait scores across all speakers and traits:

$$\text{MAE} = \frac{1}{|S| \cdot 5} \sum_{s_i \in S} \sum_{d \in D} |p_{s_i,d}^{\text{src}} - p_{s_i,d}^{\text{tgt}}| \tag{9}$$

Lower MAE indicates better absolute score preservation. While Pearson and Spearman capture rank-order consistency, MAE captures whether the actual score magnitudes are preserved.

This appendix consolidates all supplementary metric results referenced in the main text.

	Source (Ja)					Human En				
speaker	N	E	O	A	C	N	E	O	A	C
Luffy	.25	.90	.80	.70	.30	.21	.93	.85	.78	.32
Zoro	.45	.35	.40	.30	.70	.45	.35	.50	.40	.70
Nami	.34	.68	.57	.79	.89	.45	.60	.55	.65	.75
Sanji	.65	.75	.50	.40	.60	.45	.75	.50	.65	.70
Chopper	.72	.65	.58	.78	.45	.72	.45	.60	.80	.50
Robin	.34	.48	.89	.32	.85	.34	.48	.89	.72	.85
Franky	.34	.82	.68	.54	.79	.45	.75	.60	.65	.70
Brook	.45	.75	.65	.60	.55	.34	.79	.88	.82	.65
Ace	.75	.35	.40	.25	.45	.72	.45	.58	.68	.60
Jinbe	.35	.65	.55	.80	.75	.35	.60	.70	.85	.80
Hancock	.25	.30	.40	.15	.70	.72	.35	.48	.22	.65
Buggy	.78	.65	.42	.31	.48	.72	.68	.45	.32	.40
Crocodile	.65	.35	.20	.15	.70	.65	.35	.40	.20	.70
Sabo	.65	.70	.55	.45	.60	.45	.70	.60	.55	.65
Usopp	.78	.61	.59	.54	.43	.72	.65	.58	.70	.50

Table 12: Per-speaker Big Five personality vectors for One Piece (Ja source and Human En translation). Predictor: Qwen3-32B, few-shot. Scores are on a 0–1 scale.

Source (Japanese)	**Human English**	**DeepL English**	**GPT-4o English**	**Google English**
うるせえ！でかい図体(ずうたい)おろすぞ！	Can it! I'm gonna fillet that giant body of yours!	Shut up! I'm going to take you down!	Shut up! I'll take down that big body!	Noisy! I'll take down your big body!
おい… おめえ 生きてんのか？	Hey, are you alive? Where is he? What happened here?!	Hey… Are you still alive?	Hey… are you alive?	Hey… Hey, are you alive?
おい まりもヘッド 俺から離れんじゃねえぞ	Hey, Moss-Head! Don't fall behind.	Hey, Marimo Head, don't you dare leave my side.	Hey, Marimo Head, don't you dare leave me.	Hey Marimo Head, don't leave me.

Table 13: Three representative Sanji source–translation utterance pairs, illustrating the system-dependent handling of his rough register (*urusee*, *ore*, imperative forms). Human English is the OpenSubtitles human-reference subtitle; DeepL, GPT-4o, and Google English are the corresponding utterance-level MT outputs. The first row shows the cleanest contrast: DeepL and GPT-4o keep the rough command-like force with "Shut up!" while Google produces the descriptive "Noisy!".

Repeated-Utterance-Sampling Stability. The intervals in Table 2 keep the utterance sample fixed. To measure the uncertainty from the utterance sample itself, we re-drew the input of every character 20 times (random samples of at most 100 utterances per character, the same cap as in the main experiments) on *One Piece* Ja→En and recomputed personality consistency each time. The standard deviation across samples is about 0.04 to 0.05 for every system (Table 14). GPT-4o changes most: its mean falls from 0.873 to 0.814, level with DeepL (0.820) and Human (0.813), while the other systems move by 0.03 or less. The sampling intervals of all six systems overlap, and thus the system rankings in Table 2 are uncertain.

Per-Corpus Results: Pearson and MAE. Table 2 reports Spearman consistency for the two corpora. Table 15 gives the Pearson ($\bar{r}$) and MAE counterparts under the same scope (*Game of Thrones*: 11 characters, mean over all 12 language pairs (DeepL covers 10); *One Piece*: Ja→En, 15 characters). On *Game of Thrones*, all three metrics agree on the extreme rankings: Qwen3 leads and Human trails under all of $\bar{\rho}$, $\bar{r}$, and MAE; on *One Piece*, GPT-4o leads and Qwen3 trails. Middle ranks reorder by at most one position, and the largest displacement appears under MAE because it penalises absolute deviation rather than rank. The supplementary metrics thus corroborate the main-text Spearman conclusions.

J Trait-Level Patterns Across Languages

This appendix examines the per-trait composition of the cross-lingual personality consistency results in §5.1. We report a language-level per-trait breakdown for *Game of Thrones*, a descriptive target-language grouping for the same corpus, and a per-trait summary for *One Piece*.

System	Main	Resampled	95% interval
GPT-4o	0.873	0.814	[0.718, 0.910]
Aya-expanse	0.830	0.799	[0.713, 0.886]
Human	0.803	0.813	[0.739, 0.887]
DeepL	0.790	0.820	[0.714, 0.925]
Google	0.757	0.786	[0.712, 0.860]
Qwen3	0.738	0.747	[0.654, 0.840]

Table 14: Repeated-utterance-sampling stability on *One Piece* Ja→En (15 characters; Qwen3-32B predictor, few-shot). *Main* is the point estimate of Table 2, obtained from a single fixed utterance sample; *Resampled* is the mean over 20 independent random samples of at most 100 utterances per character, and the interval is that mean ± 1.96 standard deviations. Rows are ordered by the main-table estimate; the resampled ordering differs.

	Human	Google	DeepL	GPT-4o	Qwen3	Aya-expanse
Game of Thrones (11 chars, 12-lang mean)						
$\bar{r}$	0.810	0.842	0.843	0.849	**0.861**	0.848
MAE	0.237	0.199	0.212	0.193	**0.192**	0.200
One Piece (Ja→En, 15 chars)						
$\bar{r}$	0.821	0.749	0.803	0.859	0.704	0.828
MAE	0.341	0.383	0.319	0.271	0.427	0.279

Table 15: Pearson ($\bar{r}$) and MAE per translation system for the two corpora, the supplementary-metric counterpart to the Spearman main results (Table 2). *Game of Thrones*: mean over all 12 language pairs (DeepL covers 10), 11 characters; *One Piece*: Ja→En, 15 characters. Predictor: Qwen3-32B, few-shot.

J.1 Game of Thrones (English to 12 Target Languages)

To examine whether the mean consistency scores in Table 2 obscure trait-specific vulnerabilities, we disaggregate *Game of Thrones* consistency by Big Five dimension. Averaged over the six translation systems and 12 target languages, EXTRAVERSION is the least rank-preserved trait ($\rho = 0.721$), whereas OPENNESS is the most stable ($\rho = 0.944$); the per-language breakdown below shows where this gap concentrates.

Table 16 reports the same data sliced by target language. The lowest-preserved trait per language takes only two values: EXTRAVERSION in 8 of 12 languages and CONSCIENTIOUSNESS in the remaining 4; OPENNESS, AGREEABLENESS, and NEUROTICISM are never the lowest in any language. The EXTRAVERSION shortfall is not driven by a single weak system: only 3 of 70 GoT system–language cells fall below $\rho = 0.5$ (Qwen3-de 0.167, Google-es 0.388, GPT-4o-pt 0.414), and the per-system range within each language for EXTRAVERSION is typically narrow (e.g., 0.144 for Arabic, 0.277 for Spanish). We treat the E/C split as two empirical regimes in our 12-language sample, not a typological pattern.

Lang	O	C	E	A	N	$\bar{\rho}$	lowest
ar	0.976	0.842	0.735	0.889	0.919	0.872	E
de	0.912	0.759	0.580	0.929	0.954	0.827	E
es	0.966	0.883	0.521	0.947	0.846	0.833	E
fr	0.944	0.926	0.805	0.828	0.847	0.870	E
hi	0.903	0.467	0.723	0.747	0.785	0.725	C
ja	0.958	0.871	0.662	0.907	0.808	0.841	E
ko	0.965	0.763	0.709	0.913	0.960	0.862	E
pt	0.944	0.910	0.736	0.877	0.851	0.864	E
ru	0.921	0.593	0.803	0.925	0.748	0.798	C
tr	0.981	0.747	0.794	0.869	0.850	0.848	C
vi	0.961	0.841	0.889	0.899	0.972	0.912	C
zh	0.893	0.775	0.726	0.907	0.832	0.827	E
mean	0.944	0.785	0.721	0.888	0.864	—	—

Table 16: Per-trait personality consistency on *Game of Thrones* En→X, by target language, averaged over the six translation systems (DeepL omitted for Hindi and Vietnamese). The *lowest* column lists the trait with the smallest Spearman in that language.

Group	O	C	E	A	N	Lowest ρ	Max $\lvert\Delta\rvert$
East Asian	0.938	0.803	0.699	0.909	0.867	E	A
Western European	0.942	0.870	0.660	0.895	0.875	E	E
Other	0.948	0.698	0.789	0.866	0.855	C	C

Table 17: Descriptive target-language grouping of the 12 *Game of Thrones* target languages. East Asian: ja, ko, zh. Western European: de, es, fr, pt. Other (residual set, not a region): ar, hi, ru, tr, vi. Each numeric cell is the mean Spearman ρ over the six translation systems and the languages in that group. *Lowest* ρ: trait with the smallest within-group mean Spearman. *Max* $|\Delta|$: trait with the largest within-group mean absolute trait-score shift.

Target-Region Grouping. Table 17 groups the 12 *Game of Thrones* target languages into three target-language groups. East Asian and Western European targets share the same rank-preservation pattern: EXTRAVERSION is the least rank-preserved trait in both groups (mean $\rho = 0.699$ and 0.660, respectively). Their absolute-shift patterns differ. The two rightmost columns of Table 17 capture different notions of change: *Lowest* ρ indicates rank-order disruption across characters, whereas *Max* $|\Delta|$ indicates absolute score shift; the two need not select the same trait. East Asian targets show the largest group-mean $|\Delta|$ on AGREEABLENESS, and this is unanimous within the group: ja, ko, and zh all show A as the maximum-shift trait. Western European targets have the largest group-mean $|\Delta|$ on EXTRAVERSION, but this is driven mainly by German and Spanish; French shows A as the maximum-shift trait and Portuguese shows C. The remaining five languages (Arabic, Hindi,

Trait	Mean ρ	Mean $\lvert\Delta\rvert$
O	0.804	0.343
C	0.827	0.266
E	0.876	0.227
A	**0.727**	**0.488**
N	0.759	0.360

Table 18: Per-trait mean Spearman (ρ) and mean absolute trait-score shift ($|\Delta|$) on *One Piece* Ja→En, averaged over the six translation systems. A (bold) is both the least rank-preserved and the most shifted trait.

Russian, Turkish, Vietnamese) show CONSCIENTIOUSNESS as the lowest-ρ trait and, narrowly, as the largest group-mean $|\Delta|$ trait, largely driven by Hindi ($\rho = 0.467$, $|\Delta| = 0.370$). These groupings are descriptive over our 12-language subtitle sample; they suggest possible target-group variation, but are not sufficient to establish a typological or cultural mechanism.

J.2 One Piece (Japanese to English)

Across the six Ja→En systems on *One Piece*, AGREEABLENESS is the least rank-preserved trait (mean $\rho = 0.727$) and also has the largest absolute trait-score shift (mean $|\Delta| = 0.488$); the remaining traits range from $\rho = 0.759$ for NEUROTICISM to $\rho = 0.876$ for EXTRAVERSION (Table 18). This pattern aligns with the case study in Appendix D, where Japanese honorifics and sentence-final particles carry agreeableness cues that lack direct English equivalents.

K Few-Shot vs. Zero-Shot Evaluation

Table 2 reports the main results using few-shot prompting. We also re-run all system combinations in a zero-shot setting (same prompt, no examples) under both predictors (Qwen3-32B and GPT-4o); Table 19 compares few-shot and zero-shot on *Game of Thrones* and *One Piece*.

Few-shot vs. zero-shot results for both corpora are reported in Table 19.

Summary. Few-shot vs. zero-shot effects vary by predictor and corpus, but the qualitative conclusions are unchanged: no single system consistently dominates, and the main findings hold regardless of which predictor or prompt format is used.

L Prompt Language Analysis

All main *Game of Thrones* and *One Piece* experiments use a Chinese-language evaluation prompt (§4). We choose this protocol based on a prompt-language diagnostic (Chinese vs. English evaluation prompt) on the *One Piece* Ja→En direction, for which English-prompt runs are available. This diagnostic justifies the main prompt design rather than re-ranking systems.

Table 20 compares Chinese vs. English evaluation prompts on the *One Piece* Ja→En direction for both predictors. We adopt the Chinese prompt as a diagnostic-motivated design choice: a non-English instruction language is intended to reduce the salience of the predictor's English-centric pretraining priors and encourage reliance on the input text.

Summary. The Chinese evaluation prompt is preferable for two reasons: (1) it is intended to reduce the predictor's tendency to bypass the input text in favor of English-centric pretraining priors; and (2) it aligns with Qwen3-32B's stronger Chinese-language capabilities, which may yield more discriminative personality predictions. We therefore use the Chinese prompt for the main evaluations and report prompt-language sensitivity as a predictor diagnostic.

M Ground-Truth Values

For reproducibility, Table 21 lists the Big Five ground-truth values for all evaluated corpora. We also show the independent Fandom-wiki source where available (0–1 scale; N E O A C).

N External Predictor Validation on Acted Dialogue

Our fandom-derived labels validate the ground truth, but not the predictor itself. To check that the predictor's scores track personality impressions that humans actually form from dialogue, we validate the paper's predictor prompt (few-shot, identical configuration to the main experiments) against PA-IEMOCAP (Gao et al., 2025), which provides human observer Big Five annotations for acted English dialogue: 302 dialogue-level profiles from 151 conversations performed by 10 actors. Unlike our main corpora, the enacted speakers are not famous characters, and thus the predictors have no pretrained knowledge of them; agreement here can only come from reading the dialogue.

From anonymized transcripts alone, both predictors reach mean Spearman correlations of 0.46 to 0.59 with the human annotations (Table 22), with a consistent trait pattern across predictors

Corpus	Pred.	Setting	Human	Google	DeepL	GPT-4o	Qwen3	Aya-expanse
GoT (12-lang, 11 characters)	Qwen3	FS	0.808	0.842	0.841	0.846	**0.861**	0.845
		ZS	0.819	0.825	0.850	**0.856**	0.824	0.833
	GPT-4o	FS	0.862	0.866	**0.892**	0.873	0.867	0.869
		ZS	0.896	0.932	**0.937**	0.926	0.928	0.930
One Piece	Qwen3	FS	0.803	0.757	0.790	**0.873**	0.738	0.830
		ZS	0.746	0.656	0.767	**0.799**	0.594	0.780
	GPT-4o	FS	0.918	0.796	0.913	0.909	0.806	**0.932**
		ZS	0.802	0.852	0.838	**0.864**	0.731	0.818

Table 19: Few-shot (FS) vs. zero-shot (ZS) personality consistency ($\bar{\rho}$) for both predictors on the two corpora. *Game of Thrones*: mean over all 12 language pairs (DeepL covers 10), 11 characters; *One Piece*: Ja→En, 15 characters. Bold = best system per row.

Predictor	Prompt	Human	Google	DeepL	GPT-4o	Qwen3	Aya-expanse
GPT-4o	CN	0.918	0.796	0.913	0.909	0.806	**0.932**
	EN	0.864	0.844	0.867	**0.914**	0.811	0.894
Qwen3-32B	CN	0.803	0.757	0.790	**0.873**	0.738	0.830
	EN	0.713	0.665	**0.844**	0.769	0.777	0.722

Table 20: Personality consistency ($\bar{\rho}$, FS) under Chinese vs. English evaluation prompts on the *One Piece* (Ja→En) direction, for both predictors. Bold = best system per row.

(AGREEABLENESS and NEUROTICISM highest, CONSCIENTIOUSNESS lowest). The predictor scores therefore track the personality impressions that human observers form from dialogue, which is the construct this paper measures. All per-trait intervals exclude zero. Because the 302 profiles come in pairs from 151 conversations performed by 10 actors, we also recomputed the mean intervals with cluster bootstraps that resample whole conversations (GPT-4o $[0.54, 0.63]$, Qwen3-32B $[0.41, 0.51]$) and whole actors (GPT-4o $[0.55, 0.62]$, Qwen3-32B $[0.41, 0.50]$); both leave the intervals essentially unchanged, consistent with the annotation variance lying almost entirely at the dialogue level (95.6%) rather than the actor level. The paper's main predictor Qwen3-32B has lower external validity than GPT-4o. We retain it as the main predictor because it is less driven by pretrained character knowledge on our corpora (Appendix A), and report GPT-4o as the higher-validity cross-check throughout.

PA-IEMOCAP also isolates the mechanism behind the name conditions in Appendix A. Replacing the anonymous speaker label with the corpus speaker identifier leaves agreement essentially unchanged for both predictors: 0.582 (identifier) against 0.586 (anonymized) for GPT-4o, and 0.461 against 0.458 for Qwen3-32B. An identifier without pretrained associations contributes nothing, and thus a large Named-to-Anonymized gap on famous characters measures the name prior itself, not a general effect of naming.

	PDB (crowd votes)					Fandom-wiki (LLM)				
Character	N	E	O	A	C	N	E	O	A	C
Game of Thrones										
Arya Stark	.38	.59	.46	.41	.45	.72	.46	.84	.33	.70
Cersei Lannister	.87	.67	.21	.10	.60	.92	.62	.38	.06	.24
Daenerys Targaryen	.66	.52	.67	.54	.71	.72	.68	.56	.49	.61
Davos Seaworth	.33	.58	.58	.75	.67	.32	.58	.59	.89	.86
Jaime Lannister	.42	.75	.38	.41	.40	.54	.66	.46	.42	.49
Jon Snow	.46	.27	.39	.79	.67	.62	.28	.48	.79	.87
Petyr Baelish	.18	.62	.85	.35	.88	.28	.64	.84	.04	.88
Samwell Tarly	.67	.40	.85	.80	.44	.84	.22	.88	.91	.63
Sansa Stark	.65	.55	.31	.64	.69	.72	.46	.52	.58	.78
Tyrion Lannister	.47	.80	.89	.44	.44	.72	.62	.89	.68	.71
Tywin Lannister	.30	.52	.44	.18	.95	.18	.42	.57	.09	.94
One Piece										
Luffy	.21	.98	.85	.65	.19	.28	.92	.84	.71	.32
Zoro	.34	.31	.19	.30	.57	.34	.22	.41	.46	.92
Nami	.66	.73	.64	.47	.65	.74	.58	.52	.66	.82
Usopp	.83	.79	.80	.69	.33	.87	.64	.74	.81	.59
Sanji	.68	.73	.61	.80	.89	.42	.78	.72	.72	.81
Robin	.20	.32	.92	.61	.79	.18	.29	.93	.69	.84
Franky	.43	.93	.69	.55	.42	.42	.93	.91	.74	.60
Brook	.19	.88	.81	.88	.75	.42	.84	.78	.79	.68
Chopper	.78	.42	.68	.95	.69	.73	.49	.59	.93	.81
Ace	.66	.83	.61	.62	.22	.72	.74	.53	.66	.31
Sabo	.40	.81	.90	.92	.81	.42	.68	.88	.79	.61
Buggy	.85	.86	.60	.15	.30	.82	.88	.42	.14	.34
Hancock	.47	.62	.47	.20	.44	.75	.69	.43	.24	.70
Jinbe	.33	.42	.58	.50	1.00	.19	.51	.73	.88	.93
Crocodile	.06	.55	.81	.17	.90	.55	.68	.62	.08	.82

Table 21: Per-character Big Five ground truth (0–1): PDB (crowd votes) and an independent LLM-scored wiki source. A dash marks a source with no entry for that character. Coverage: *Game of Thrones* 11; *One Piece* 15 (both sources cover all characters in both corpora). Character-level PDB–wiki agreement for *Game of Thrones* and *One Piece* is reported in §6.1.

Predictor	**Condition**	**Mean ρ**	N	E	O	A	C
GPT-4o	Anonymized	.586 [.54,.62]	.716 [.66,.77]	.454 [.35,.55]	.569 [.49,.65]	.794 [.75,.83]	.398 [.29,.50]
GPT-4o	Identifier	.582 [.54,.62]	.733 [.68,.78]	.410 [.30,.51]	.580 [.49,.66]	.792 [.74,.83]	.395 [.29,.49]
Qwen3-32B	Anonymized	.458 [.41,.50]	.620 [.55,.68]	.296 [.18,.41]	.504 [.41,.59]	.656 [.58,.72]	.212 [.10,.32]
Qwen3-32B	Identifier	.461 [.41,.50]	.632 [.56,.69]	.268 [.15,.38]	.519 [.43,.60]	.675 [.60,.74]	.209 [.10,.32]

Table 22: External predictor validity on PA-IEMOCAP: Spearman correlation between predicted Big Five scores and human observer annotations ($n = 302$ dialogue-level profiles), overall and per trait, with 95% percentile-bootstrap confidence intervals over profiles (gray). Anonymized = anonymized transcripts; Identifier = the corpus speaker identifier included. The near-zero Anonymized–Identifier gap for both predictors shows that naming alone contributes nothing when the name carries no pretrained associations.