

Evaluating Pragmatic Appropriateness in English–Vietnamese Machine Translation with Dyadic Profiles

Zheng-Yu Huang¹ Richard Tzong-Han Tsai^{2,3*}

11430084@st.chjhs.tp.edu.tw, thtsai@g.ncu.edu.tw

¹Chingshin Academy, Taipei, Taiwan

²Department of Computer Science and Information Engineering, National Central University, Taiwan

³Center for GIS, RCHSS, Academia Sinica, Taiwan

Abstract

English workplace requests often omit relationship information needed for pragmatically appropriate Vietnamese person reference and politeness. We evaluate translations generated with and without explicit sender–recipient profiles for 30 requests using two large language models (LLMs), Gemini 3.5 Flash and SEA-LION, with judgments from two native speakers of Vietnamese. Two native speakers of Vietnamese judged whether translations were correct and natural and whether they suited the displayed relationship. Adding profiles produced no statistically detectable overall improvement in appropriateness; translations generated without profiles already received high ratings. In a prespecified role-reversal evaluation, the same translations were judged first under their intended hierarchical relationships and later under reversed relationships. All 20 Gemini translations generated with profiles were judged appropriate under the intended relationship, compared with four after reversal; SEA-LION’s result was inconclusive. A post-hoc side-by-side comparison also favored Gemini’s translations generated with profiles but found no detectable SEA-LION preference. These results suggest that when most translations receive positive yes/no ratings, role reversal and direct comparison can reveal differences those ratings alone miss. Fixed evaluation order and reuse of the annotators in the follow-up limit causal interpretation.

1 Introduction

When translating from English, the source utterance may underspecify attributes of the message sender and recipient or other contextual information needed to choose the socially and pragmatically appropriate forms in the target language (Vincent et al., 2022). In Vietnamese, the relationship between the interlocutors can affect how speakers refer

to themselves and address others through pronouns, kin terms, names, titles, or status terms (Sidnell and Shohet, 2013). These contexts can also affect politeness and deference marking through particles such as *dạ* and *ạ* (Sidnell and Shohet, 2013; Nguyen and Ho, 2013). A machine translation (MT) output can preserve what is being requested and be grammatical, yet use unsuitable forms of address, a mismatched level of formality, or an inappropriate way of making the request for that relationship. Evaluation without relationship information cannot directly assess this fit.

Vu et al. (2024) test a translation model that also predicts which source expressions refer to the same entity. Their English-to-Vietnamese evaluation targets the distinction between inclusive *we* (*chúng ta*) and exclusive *we* (*chúng tôi*), a distinction absent in English. On 173 cases, pronoun-translation accuracy was 18.50% for their coreference-aware Trans+Coref model, versus 12.71% for Base Doc and 14.45% for G-Transformer. Base Doc is their Transformer baseline trained on multiple sentences; G-Transformer is a document-level comparison model that incorporates context in its encoder. These are systems from the cited study, not our profile-free BASE condition. Even with context, most targeted pronouns were translated incorrectly.

Workplace requests are messages asking someone to carry out a work-related action. They provide useful test data for evaluating English-to-Vietnamese pragmatic appropriateness because workplace roles allow sender–recipient relationships to be defined explicitly. A Vietnamese role-play study associated relative power with request-strategy and address-term choices (Nguyen and Ho, 2013). A dyadic profile describes the message sender, recipient, and their relationship; here, the profiles represent synthetic workplace scenarios. To our knowledge, this is the first direct test of whether English-to-Vietnamese MT outputs generated from explicit workplace sender–recipient profiles fit their

*Corresponding author.

intended relationships.

We study 30 English workplace requests in three relationships: junior-to-manager (*upward*), manager-to-junior (*downward*), and equal rank (*peer*). For Gemini and SEA-LION, BASE generates a translation without sender–recipient profiles; DYAD adds the intended profiles; and SWAPPED generates a new translation after the roles are exchanged. Google Translate supplies a BASE-only conventional MT comparison.

Two native speakers of Vietnamese completed two passes. Without profiles, they judged *acceptability*: whether a translation was correct, natural, and preserved the requested meaning and action. With profiles visible, they judged *pragmatic appropriateness*: whether its wording suited the displayed relationship. Each hierarchical DYAD translation was then judged in a later task with the roles reversed, keeping the Vietnamese text unchanged. This differs from SWAPPED, which generates a new translation. In an explicitly post-hoc follow-up, the same annotators compared 58 displayed BASE–DYAD pairs side by side.

In this paper, we address the following research questions:

RQ1. Does providing explicit sender and recipient role information in the prompt improve pragmatic appropriateness in English-to-Vietnamese machine translation output?

RQ2. For hierarchical cases, how does the judged pragmatic appropriateness of a realized DYAD translation change when the same Vietnamese text is evaluated under its intended sender–recipient profiles versus the reversed profiles?

Contribution. We provide a two-stage human-evaluation protocol and a fixed-translation role-reversal test that distinguish translation acceptability from relationship-dependent fit.

Main findings. The clearest result from the pre-specified analyses was Gemini’s fixed-text contrast: 20/20 appropriate under intended profiles versus 4/20 under later reversed profiles (Holm-adjusted $p = .00006104$). The text stayed unchanged, but intended profiles were always rated first. SEA-LION’s contrast was inconclusive. RQ1 detected no overall improvement from adding profiles. Post-hoc, 41/44 choices between Gemini translations favored DYAD: 18/20 for one annotator and 23/24 for the other. These are individual preferences, not 44 independent source pairs. Neither SEA-LION test detected a preference.

2 Related Work

2.1 Vietnamese person reference and social relations

Vietnamese person reference forms an integrated relational system encompassing pronouns, kin and status terms, names, titles, and zero reference (Sidnell and Shohet, 2013; Ton, 2018). These choices depend on the speaker’s attributes and the relationship between participants, including age, status, closeness, stance, and the activity (Sidnell and Shohet, 2013; Ton, 2018). Even in a two-person conversation, how each person refers to themselves may depend on different factors from how they address the other. Equal rank does not determine a single neutral form of address: choices can express closeness, distance, or mutual respect, or temporarily downplay hierarchy (Sidnell and Shohet, 2013).

The relationship also affects how a request is phrased. In Vietnamese role-play data, equal-power speakers used imperatives more often, while lower-to-higher requests favored *query-preparatory forms*: questions about the conditions for carrying out an action, such as ability or willingness. Address terms, honorific elements, and modal particles were also common (Nguyen and Ho, 2013). An indirect request is not automatically more appropriate than a direct one; an imperative can suit the relationship when its wording and forms of address fit the context. Speakers may omit explicit person reference when context makes clear who is meant, but omission can still create ambiguity or sound rude (Ton, 2018).

These choices pose a translation problem when English leaves participant roles unstated: the Vietnamese translation must still suit the intended interaction, whether person reference is explicit or omitted. Preserving the requested action and grammaticality alone therefore does not ensure an appropriate form of address or request. The English wording may nevertheless hint at a relationship, so missing explicit roles should not be treated as an absence of relationship cues (Vincent et al., 2022).

2.2 Pragmatics and context in machine translation

MT studies use different kinds of context. Document-context models use surrounding sentences or information about expressions that refer to the same entity. In English-to-Vietnamese translation, such models can improve BLEU (Bilingual Evaluation Understudy), which measures word-

sequence overlap with reference translations, while remaining weak on distinctions such as inclusive versus exclusive *we* (Vu et al., 2024). Other work supplies speaker or addressee attributes directly, including gender, number, or desired formality. These attributes can control socially or grammatically marked wording but can also introduce unwanted changes (Vincent et al., 2022). English-to-Vietnamese neural machine translation (NMT) has likewise used formal and informal labels, with contrastive data containing person-reference and deferential markers (Vakharia et al., 2023). In these settings, however, the desired attribute is specified directly rather than inferred from a workplace relationship.

Other approaches condition translation on a speaker’s identity or on instructions about audience, purpose, and tone. Speaker-conditioned English-to-Japanese models improved some pronoun and sentence-final-particle choices but not honorific accuracy in a small manually evaluated subset; honorific choices depend on the speaker–hearer relationship, not just the speaker (Nagato and Matsuzaki, 2025). Translation guided by audience, purpose, tone, or setting can also receive human judgments that differ from general automatic metrics (Kayano and Sugawara, 2025). A recent preprint extends instruction-conditioned translation to broader audience and intent adaptation (Merx et al., 2026). Our DYAD condition instead supplies factual descriptions of both workplace participants. We distinguish generating a new translation after role reversal (SWAPPED) from judging the original DYAD translation under a reversed displayed relationship.

2.3 Human evaluation of pragmatic translation

General translation quality and fit to a specific relationship are different questions. Human document-level MT evaluation has assessed pronouns, titles, honorifics, social relationships, and style, while also finding disagreement among professional annotators (Kim, 2025). Human rankings for a specified translation purpose can differ from rankings based on a general automatic metric (Kayano and Sugawara, 2025). PhoMT is a Vietnamese–English parallel dataset and MT benchmark (Doan et al., 2021). In its human evaluation, three Vietnamese annotators compared anonymously presented, shuffled translations; the study reported agreement before discussion and then resolved disagreements. Its best-translation task did not separately assess

Source-form stratum	Up	Down	Peer	Total
Imperative/direct	3	4	3	10
Modal/query-preparatory	3	3	4	10
Elaborated/mitigated	4	3	3	10
Total	10	10	10	30

Table 1: Purposive allocation of source requests across operational source-form strata and relationship directions.

social context. Work on formality control also supports evaluating meaning preservation, fluency, and the targeted social dimension separately rather than combining them in one score (Niu et al., 2018).

3 Experimental Design

3.1 English source requests

We selected 30 workplace requests from the Enron Email Corpus (Klimt and Yang, 2004). Candidate discovery used a pinned revision of the corbt/enron-emails dataset.¹ Selection was manual and purposive. Each request had to contain one reasonably clear request head act (the clause or phrase performing the request), be comprehensible without its email, omit personal information and opaque references needed for interpretation, fit the controlled workplace relationships, and fill one of three ten-item surface-form quotas. The private source inventory records the final English texts, operational strata, and assigned relationship directions, but not record keys, source spans, checksums, or the intermediate candidate pool. Candidate discovery and source-span verification cannot be replayed from it.

The requests were grouped by their surface form: ten were imperative or direct, ten used modal or query-preparatory forms, and ten were more elaborated or heavily mitigated. These groups were used to balance the dataset; they are not independently validated pragmatic-strategy labels. We assigned ten requests to each relationship direction and kept the three groups approximately balanced across directions (Table 1). These directions were created for the experiment and do not represent the relationships between the original Enron senders and recipients. Assigning synthetic relationships does not establish that the English requests lack implicit relationship cues.

¹<https://huggingface.co/datasets/corbt/enron-emails/tree/cfc06c758093d90993abce1a43668fb7357258a6>

3.2 Social profiles and relationship directions

The participant profiles were constructed for the experiment rather than recovered from the original emails. In every scenario, both participants were described as 30-year-old Vietnamese women employed by the same company. They knew each other professionally but were not relatives or personally close, and their interactions were ordinary workplace exchanges. Holding age, gender, employer, and familiarity constant made organizational role the primary profile difference. The experiment therefore does not test the effects of age, gender, or intimacy.

The relationship scenarios were formed as follows: an upward relationship designated a junior analyst as sender and a department manager as recipient. The downward relationship reversed these roles, with a department manager sending a request to a junior analyst. The peer relationship paired an analyst in Team A with an analyst in Team B at the same organizational level. The sample comprised ten upward, ten downward, and ten peer source scenarios.

In the DYAD condition, the model received the intended sender and recipient profiles. In the SWAPPED condition, the two workplace roles were exchanged and the model generated a new translation of the same English request based on the modified profiles. For hierarchical scenarios, this reversed the relationship direction; for peer scenarios, it exchanged the team labels without changing rank. The latter served only as a stability control. All other profile fields remained unchanged.

3.3 Systems and prompting conditions

We evaluated two LLMs and one conventional MT baseline; all 210 selected outputs were generated on July 20, 2026 (UTC). Gemini used the v1beta generateContent API with identifier gemini-3.5-flash, pinned catalog version 3.5-flash-05-2026, temperature 0, and a 1,024-token limit. SEA-LION ran locally through Ollama as aisingapore/Qwen-SEA-LION-v4-32B-IT:latest, with temperature 0, num_predict=1024, num_ctx=4096, seed 42, and streaming disabled. A July 29 registry lookup described the mutable tag as a Qwen3 model with 32.8 billion parameters, stored in GGUF, a model-weight file format, with the Q4_K_M quantization setting. Because the lookup occurred after generation, it does not verify the checkpoint used at execution. Google Trans-

late used googletrans 4.0.2 against the recorded translate.googleapis.com endpoint for en-vi and exposed no decoding controls. Both LLMs produced BASE, DYAD, and SWAPPED outputs; Google Translate produced BASE only.

The preserved prompt templates and request records show what information was sent in the user and system messages. Gemini BASE used one translation-only user instruction plus the source; DYAD and SWAPPED added the complete speaker, addressee, and directed-relationship profiles to that message, with no system message. SEA-LION used a professional-translator system message and the source as the user message, adding the complete social context to the system message for DYAD and SWAPPED. Google Translate received the byte-identical source with en/vi language codes and no social context. Every rendered prompt is retained in the private records.

Automatic SDK/client retries were disabled. The runner selected the first fully validated output, preserved failures, and reused validated outputs on re-run; invalid generations were rejected, not repaired. The default allowed at most five attempts per run with exponential backoff, but attempts accumulated across runs. Invocation commands and attempt overrides were not recorded. Source bytes and profiles remained fixed, and selected provider text was retained verbatim without manual editing.

3.4 Output inventory

The inventory comprised 210 generated records: 90 per LLM (30 in each condition) and 30 Google Translate BASE records. The 60 SWAPPED records are included in this total.

Each annotator evaluated 210 outputs plus 15 hidden repeats in each of two main passes: 225 items per pass and 450 presentations, each eliciting one judgment. The 40 fixed-text counterfactual presentations brought the primary-experiment total to 490 judgments per annotator. The post-hoc preference task added 58 displayed pairs, bringing the total to 548 judgments per annotator; two additional identical pairs were structural ties and were not presented.

4 Human Evaluation

4.1 Annotators and adjudication

Two native speakers of Vietnamese, identified as A1 and A2, independently annotated every item. Both completed secondary schooling in Vietnamese,

held bachelor’s degrees in computing-related fields, had Vietnamese-NLP experience through coursework, research, or development, and met at least B1 English on the Common European Framework of Reference for Languages (CEFR); neither reported formal linguistics training. Recruitment used personal and professional CS/AI/NLP networks; B1 English was required and Vietnamese-NLP experience preferred. The main-task agreement estimated approximately 5–10 hours in flexible 45–60-minute sessions. Each primary annotator received USD 110 for the main task; follow-up compensation was set at an additional USD 20 for 58 comparisons (USD 130 in total per annotator). A Vietnamese-speaking NLP researcher served as the expert adjudicator (ADJ) but did not complete the full task. ADJ reviewed all disagreements and all CANNOT ASSESS or TECHNICAL PROBLEM responses. For each reviewed judgment, ADJ saw both annotators’ original responses, any reasons, the English source, the Vietnamese translation, and the displayed relationship; provider, model, and condition remained hidden. Comparable recent work used one native-speaker annotator per target language (Merx et al., 2026); here, two independent primary annotators covered every item, enabling agreement and rater-specific sensitivity analyses.

ADJ assigned a separate final outcome and rationale to each flagged acceptability or pragmatic appropriateness judgment, retaining both annotators’ original responses. Reliability used the original responses; the research-question analyses used the final outcomes.

4.2 Main two-stage task

The main task was divided into two complete passes over the translation set. In the first pass, annotators evaluated every English source request and Vietnamese translation without seeing the system, generation condition, or sender–recipient profiles. They judged whether the Vietnamese was a correct and natural translation that preserved the main meaning and requested action. We refer to this first-pass judgment as *acceptability*. A No response could therefore indicate either a meaning error or unacceptable Vietnamese wording; the combined question did not distinguish between these causes. The entire first pass was submitted and locked before the second pass began.

In the second pass, the same English requests and Vietnamese translations were presented again, this time with the complete sender and recipient

Task	Substantive responses	Special responses
Main acceptability and pragmatic appropriateness; counterfactual appropriateness	YES; NO	CANNOT ASSESS; TECHNICAL PROBLEM
Post-hoc preference	A better; B better; no preference	CANNOT ASSESS; TECHNICAL PROBLEM

Table 2: Response options as displayed to annotators. A/B are blinded translation labels, mapped to BASE/DYAD for analysis.

profiles. Annotators then judged whether each translation was socially and pragmatically appropriate for the displayed workplace relationship. For SWAPPED outputs, the displayed profiles were the exchanged profiles supplied at generation. Thus, context was absent throughout the first pass and available throughout the second; the two judgments were not elicited sequentially within a single presentation.

Table 2 lists the available responses for each task. In both main passes, CANNOT ASSESS and TECHNICAL PROBLEM required a short reason, whereas reasons for YES and NO were optional. We treated CANNOT ASSESS as missingness rather than as a midpoint or negative judgment. In our terminology, a *main-task item* is one source–translation pair scheduled in both passes, a *presentation* is one display of that item in one pass, and a *judgment* is one response to the evaluation question asked in that pass.

The deployed guides instructed annotators to use native-speaker judgment and did not impose automatic accept/reject rules for neutral wording, omitted address or person reference, ellipsis, imperatives, or overt hierarchy marking.

4.3 Fixed-translation counterfactual evaluation

For each of the 20 hierarchical sources, the task reused one DYAD translation from each LLM. Each annotator judged these 40 translations under reversed roles, without changing the Vietnamese text. Peer cases were excluded because exchanging equal-rank team labels does not reverse a hierarchy.

Intended-profile ratings came from the main appropriateness pass; reversed-profile ratings came from a later task. Intended ratings always came first, with at least 20 intervening presentations between

paired ratings. Because displayed profiles and evaluation timing changed together, memory, learning, fatigue, threshold drift, and demand effects remain possible. BASE outputs were not reversed.

Unlike SWAPPED, this task generated no new translations. It cannot establish the effect of profile prompting on generation or reveal the model’s reasoning.

4.4 Randomization and quality control

The provider, model, and generation condition remained hidden throughout annotation and adjudication. A1 and A2 received independently randomized item orders; within each annotator, both main passes used the same 225-item order. The schedule also placed at least 20 intervening items between original-role and SWAPPED outputs of the same source and between an original item and its hidden repeat. Fifteen scheduled items were duplicated without identification; each appeared in both passes, producing 15 original-repeat comparisons per construct and annotator. We define within-annotator consistency as the exact-category match rate on these original-repeat pairs.

5 Statistical Analysis

5.1 Analysis scope and inferential approach

Our paired analyses concern these 30 selected requests, the generated translations, and the annotation procedure used here, not broader populations of languages, speakers, or model outputs. RQ1 and RQ2 use the prespecified exact McNemar test at $\alpha = .05$. Source-pair percentile-bootstrap intervals (Efron, 1979) describe effect magnitude; the exact tests determine statistical significance. Non-significance is not evidence of equivalence or no effect.

As prespecified, we repeated RQ1 and RQ2 separately using each annotator’s original judgments, with the same exact tests and Holm families. These sensitivity checks assess dependence on adjudication without pooling the two annotators’ responses.

5.2 Reliability

Inter-annotator reliability was calculated before adjudication and separately for acceptability, pragmatic appropriateness, and counterfactual pragmatic appropriateness; the three judgment sets were not pooled. Hidden repeats were excluded from the inter-annotator calculations. For each judgment

set, we report raw agreement across all four response categories, raw agreement and Cohen’s κ (Cohen, 1960) on the YES/NO subset, and, on the same YES/NO subset, prevalence- and bias-adjusted kappa, $PABAK = 2p_o - 1$. Because κ is sensitive to response prevalence, it is interpreted alongside raw agreement and PABAK (Byrt et al., 1993), with a prevalence warning when the pooled proportion of YES responses is at least .90.

We obtained 95% confidence intervals (CIs) for raw agreement, κ , and PABAK from 2,000 item bootstraps (seed 20260718). For each statistic, bootstrap samples in which it was undefined were excluded from its interval calculation and counted. The interval limits were the 2.5th and 97.5th percentiles of the remaining estimates. Within-annotator consistency used the 15 hidden original-repeat pairs for each main construct. The prespecified rule required $PABAK \geq .60$ and raw YES/NO agreement $\geq .80$ for both main constructs. Because $PABAK = 2p_o - 1$, these thresholds are equivalent on the same binary pairs; both are retained. No counterfactual threshold was prespecified.

5.3 RQ1: BASE-DYAD comparison

For each LLM separately, RQ1 compares BASE and DYAD outcomes for the same source requests, giving at most 30 paired sources per comparison. The source request is the paired unit, and the analysis uses post-adjudication outcomes. We report three outcomes: acceptability, pragmatic appropriateness, and joint success, which requires both judgments to be YES. Google Translate is reported descriptively and is not part of this contrast because it has no DYAD condition.

For each outcome, we use an exact two-sided McNemar test (McNemar, 1947) on improved and worsened discordant pairs. The implementation doubles the smaller binomial tail under $B \sim \text{Binomial}(b + c, .5)$, caps the result at one, and returns $p = 1$ when $b + c = 0$. We report the paired success-rate difference, DYAD minus BASE, with a descriptive source-pair percentile-bootstrap 95% interval. Holm’s step-down procedure (Holm, 1979) corrects the two LLM comparisons within each outcome family.

5.4 RQ2: Fixed-text fit comparison

The primary RQ2 analysis uses 20 hierarchical sources per LLM. Each pair contains two final appropriateness judgments of the same DYAD translation: one from the earlier main pass under the

intended profile and one from the later role-reversal task under the reversed profile. The paired difference is intended minus reversed. We use the same exact McNemar implementation, descriptive percentile-bootstrap interval, and two-LLM Holm family as above.

The newly generated SWAPPED outputs provide a separate descriptive check. We report their appropriateness under the exchanged profiles and whether each DYAD–SWAPPED pair contains identical text after normalization. For this text-identity check, Unicode Normalization Form C (NFC) gives canonically equivalent character sequences the same form; we then apply case folding, collapse whitespace, and remove terminal punctuation. For each LLM, we also compare the final appropriateness outcomes of the ten peer DYAD translations with their SWAPPED counterparts. This checks stability when equal-rank team labels are exchanged. These peer pairs receive no inferential test and are not pooled with the hierarchical RQ2 comparison. Outcomes by source-form group are also descriptive because each group contains only ten sources.

5.5 Special responses and sensitivity analysis

CANNOT ASSESS and TECHNICAL PROBLEM remain distinct. For acceptability and pragmatic appropriateness, the primary analyses exclude any comparison pair containing either response and report the number excluded.

For joint success, we first combine the two judgments for each translation. A TECHNICAL PROBLEM response takes precedence and excludes the affected comparison pair. Otherwise, a No on either question means joint failure, even if the other response is CANNOT ASSESS. With no No response, any CANNOT ASSESS means a missing joint outcome; two YES responses mean joint success. A comparison pair containing a missing joint outcome is excluded.

The prespecified sensitivity analysis instead treats CANNOT ASSESS as No; TECHNICAL PROBLEM remains excluded. Holm correction is applied separately within the primary and sensitivity-analysis families. Success-rate denominators contain only YES and No outcomes, and special responses are reported separately by system and condition.

System	Condition	Acceptable	Appropriate
Gemini	BASE	30/30	28/30
	DYAD	30/30	30/30
	SWAPPED	30/30	30/30
SEA-LION	BASE	27/30	25/29
	DYAD	29/30	27/30
	SWAPPED	25/30	25/30
Google Translate	BASE	22/30	25/29

Table 3: Post-adjudication success counts. Denominators exclude CANNOT ASSESS; the two exclusions are described in the text.

6 Results

6.1 Descriptive outcomes and measurement quality

Table 3 reports the post-adjudication outcomes. All 210 generated records were evaluated. Before adjudication, neither annotator selected CANNOT ASSESS; one Google Translate BASE pragmatic-appropriateness item received a TECHNICAL PROBLEM response (see Table 2 for response options). After adjudication, two final outcomes were CANNOT ASSESS, one Google Translate BASE and one SEA-LION BASE, both for pragmatic appropriateness and each arising from a YES/NO disagreement that ADJ judged not assessable; no final outcome was TECHNICAL PROBLEM. The SEA-LION BASE–DYAD pragmatic-appropriateness and joint-success comparisons consequently contained 29 paired sources, and every other RQ1 comparison contained 30. Recoding CANNOT ASSESS as failure in the prespecified sensitivity analysis did not change any inferential conclusion.

Before adjudication, raw agreement was 91.4% for acceptability, 87.1% for pragmatic appropriateness, and 80.0% for counterfactual appropriateness (Table 4). For acceptability, 95.2% of the pooled responses were YES. With YES so common, expected agreement in the kappa calculation was already 90.7%. Observed agreement was only slightly higher, at 91.4%, yielding $\kappa = .072$ despite high raw agreement. This prevalence problem (Byrt et al., 1993) is why we report kappa alongside 91.4% raw agreement and PABAK = .829. Both main constructs met the prespecified reliability rule. Counterfactual PABAK was .600 [.300, .850]; its interval was wide, and no threshold was prespecified for that judgment set.

Both annotators exactly reproduced 15/15 acceptability and 14/15 pragmatic hidden-repeat judgments. The primary experiment produced 980 re-

Construct	$n_{\text{all}}/n_{\text{YN}}$	Raw agr. [95% CI]	κ [95% CI]	PABAK [95% CI]
Accept.	210/210	.914 [.876,.948]	.072 [-.042,.274]	.829 [.752,.905]
Prag.	210/209	.871 [.828,.919]	.370 [.180,.550]	.751 [.656,.837]
Counterf.	40/40	.800 [.675,.925]	.578 [.293,.800]	.600 [.300,.850]

Table 4: Pre-adjudication reliability; hidden repeats excluded. Upper row: estimate; lower row: 95% CI. Raw agreement uses four categories; κ and PABAK use YES/No. Acceptability κ reflects 95.2% pooled YES prevalence.

Judgment	A1	A2	Neither
Acceptability	3	15	0
Pragmatic appropriateness	15	10	2
Counterfactual appropriateness	6	2	0
All reviewed cases	24	27	2

Table 5: Which annotator’s response matched the final outcome in the 53 adjudicator-reviewed cases. Neither denotes a final outcome matching neither original response.

sponses, including 60 hidden-repeat responses (15 items $\times$ two constructs $\times$ two annotators). The remaining 920 responses formed 460 pairs of ratings, one rating from each annotator. The annotators agreed on 407 pairs. The other 53 required adjudication: 52 involved YES/No disagreements and one involved a special response. Table 5 shows whose response matched the final outcome.

6.2 RQ1 result: No aggregate gain

Adding dyadic profiles did not yield a statistically detectable aggregate improvement for either LLM (Table 6). Gemini’s pragmatic-appropriateness rate increased from 28/30 (93.3%) under BASE to 30/30 (100%), a paired difference of 6.7 percentage points (95% CI [0.0, 16.7]); the two discordant pairs improved and none worsened (exact $p = .500$, Holm-adjusted $p = 1.000$). For SEA-LION, the matched rate increased from 25/29 (86.2%) to 26/29 (89.7%), a difference of 3.4 points (95% CI [-10.3, 17.2]); three pairs improved and two worsened (exact and Holm-adjusted $p = 1.000$).

Acceptability and joint success followed the same pattern; no RQ1 outcome family produced an adjusted p -value below .05. The positive point estimates remain imprecise, so RQ1 supports neither an aggregate gain nor equivalence.

Outcome	n	B%	D%	Δ pp [95% CI]	Raw p b/c Holm p
<i>Gemini</i>					
Accept.	30	100	100	0.0 [0.0,0.0]	0/0 1.000
Prag.	30	93.3	100	6.7 [0.0,16.7]	2/0 .500
Joint	30	93.3	100	6.7 [0.0,16.7]	2/0 .500
<i>SEA-LION</i>					
Accept.	30	90.0	96.7	6.7 [0.0,16.7]	2/0 .500
Prag.	29	86.2	89.7	3.4 [-10.3,17.2]	3/2 1.000
Joint	29	79.3	86.2	6.9 [-10.3,24.1]	4/2 .6875

Table 6: Paired RQ1 comparisons. B/D: BASE/DYAD percentages; $\Delta = D - B$; b/c : improved/worsened. CIs: descriptive percentile bootstraps. Exact McNemar p -values: upper row, raw; lower row, Holm-adjusted within outcome.

6.3 RQ2: Gemini fixed-text contrast

Under the sequential fixed-text evaluation, Gemini’s realized DYAD outputs showed a large intended-versus-reversed fit contrast (Table 7). All 20 were appropriate under their intended hierarchical profiles, whereas 4/20 remained appropriate when the same text was later evaluated under reversed profiles. The paired difference was 80 percentage points (descriptive 95% bootstrap interval [60, 95]); 16 pairs were intended-only and none reversed-only (exact $p = .00003052$, Holm-adjusted $p = .00006104$).

For SEA-LION, 17/20 translations were appropriate under intended profiles and 13/20 under later reversed profiles, a difference of 20 points (descriptive 95% bootstrap interval [5, 40]). Four paired differences were positive and sixteen were zero; an all-zero bootstrap resample has probability $(16/20)^{20} = .0115$, so the descriptive percentile interval begins at 5 points. The prespecified exact McNemar test was nonsignificant (raw and Holm-adjusted $p = .125$) and controls the conclusion; the SEA-LION evidence remains inconclusive.

Pre-adjudication rater-specific checks retained the conclusions: no RQ1 contrast was Holm-significant (minimum adjusted $p = .500$ for A1 and $p = .250$ for A2); RQ2 remained detectable for Gemini ($p = .00006104/.00001526$) but inconclusive for SEA-LION ($p = .375/.2266$) for A1/A2, respectively.

Measure	Gemini	SEA-LION
Paired sources n	20	20
Appropriate, intended	20/20	17/20
Appropriate, reversed	4/20	13/20
Difference (pp)	80	20
95% CI (pp)	[60,95]	[5,40]
Intended-only	16	4
Reversed-only	0	0
Raw p	.00003052	.125
Holm p	.00006104	.125

Table 7: Sequential fixed-text RQ2 comparison. Intended judgments preceded reversed judgments; BASE was not reversed. Differences are intended minus reversed; intended-only and reversed-only count discordant pairs. CIs are descriptive percentile bootstraps; exact McNemar tests determine inference.

6.4 Secondary diagnostics

Under their exchanged profiles, 20/20 hierarchical SWAPPED outputs from Gemini and 15/20 from SEA-LION were pragmatically appropriate. The two normalized texts differed in every hierarchical DYAD–SWAPPED pair for both systems. Among peer pairs, they differed in 6/10 Gemini pairs and 4/10 SEA-LION pairs; nevertheless, all ten peer pairs for each system retained the same positive pragmatic outcome across conditions. In the supplementary BASE–DYAD identity check, the normalized texts differed in all 30 Gemini pairs and 28/30 SEA-LION pairs; the two identical SEA-LION pairs were peer scenarios. These diagnostics are descriptive: textual difference is neither sufficient evidence of pragmatic adaptation nor a substitute for the fixed-text RQ2 contrast.

Outcomes by source-form stratum were also descriptive. Gemini’s two BASE pragmatic failures both occurred among elaborated or heavily mitigated requests (8/10 BASE versus 10/10 DYAD); its direct and modal strata were 10/10 in both conditions. SEA-LION’s BASE and DYAD counts were 8/10 versus 9/10 for direct requests, 9/10 versus 10/10 for modal requests, and 8/9 versus 8/10 for elaborated requests. These groups are too small to support inferential claims about source form.

7 Post-Hoc Preference Experiment

7.1 Design and analysis

This post-hoc exploratory follow-up was designed after the frozen binary analysis. It tested whether direct comparison could reveal relative differences not distinguished by the high binary appropriateness rates. It complements rather than replaces the

main-study results.

Materials. The experiment reused the exact frozen BASE and DYAD outputs; no translations were regenerated. The comparison set contained $30 \times 2 = 60$ source–LLM pairs. Two SEA-LION pairs were identical after the normalization described in Section 5 and were recorded as structural ties without human presentation, leaving 58 displayed pairs per annotator. Google Translate, SWAPPED outputs, and counterfactual items were excluded.

Participants. The same two annotators, A1 and A2, independently completed the follow-up. Both had previously judged the outputs in the main task, creating the possibility of repeated-exposure effects. Their preference judgments were retained without adjudication.

Procedure. Each item presented the English source request, the intended sender–recipient profiles, and two translations labeled A and B. The mapping from A and B to BASE and DYAD was independently randomized for each annotator and remained hidden, as did the LLM identity. A/B position assignment was generated deterministically with seed 2026072901; presentation order used seed 2026072918 for A1 and 2026072927 for A2. Annotators selected one of the five preference responses in Table 2. A written reason was collected with each response but was not included in the quantitative analysis.

Analysis. We analyzed each annotator and LLM combination separately, giving four comparisons. For each, an exact two-sided binomial test compared the number of preferences for BASE with the number for DYAD, excluding no-preference and special responses. We report the DYAD proportion among these preferences, a 95% Clopper–Pearson confidence interval (Clopper and Pearson, 1934), and a Holm-adjusted p -value across the four tests. Agreement was calculated for BASE, DYAD, and no preference, excluding the displayed pair containing a special response. The two annotators’ judgments were not pooled as 116 independent observations.

7.2 Preference results

Table 8 reports the annotator-specific comparisons. Both annotators’ Gemini results favored DYAD after Holm correction; neither SEA-LION result did. For the special responses listed in Table 2, both annotators recorded CANNOT ASSESS for the same

Stratum	D/B/NP/CA	D/dir.	95% CI	Holm p
Gemini A1	18/2/10/0	18/20	[.683,.988]	.0012
Gemini A2	23/1/6/0	23/24	[.789,.999]	< .0001
SEA-LION A1	8/7/12/1	8/15	[.266,.787]	1.000
SEA-LION A2	11/3/13/1	11/14	[.492,.953]	.1147

Table 8: Post-hoc BASE–DYAD preferences. D/B/NP/CA are counts for DYAD, BASE, no preference, and CAN NOT ASSESS; D/dir. is DYAD/directional. CIs are exact Clopper–Pearson.

SEA-LION pair, and neither recorded a TECHNICAL PROBLEM.

7.3 Agreement and pair-level outcomes

A1 and A2 matched on 35/57 eligible pairs (61.4%) for the three substantive categories; one special-response pair was excluded. Across all five categories, they matched on 36/58 pairs (62.1%). In 19 of the 22 disagreements, one annotator chose BASE or DYAD while the other chose no preference; in the other three, one chose BASE and the other DYAD. None was adjudicated.

Across all 60 source–LLM pairs, the pair-level outcome breakdown was 21 unanimous DYAD, 3 unanimous BASE, 11 unanimous no preference, 19 mixed-with-no-preference, 3 direct splits, 1 unavailable special-response pair, and 2 structural ties. The module was explicitly post-hoc, was completed by the original annotators, and is interpreted separately from primary RQ1.

8 Discussion

Gemini’s DYAD translations were appropriate under 20/20 intended and 4/20 later reversed profiles (Holm-adjusted $p = .00006104$); SEA-LION’s 17/20 versus 13/20 contrast was inconclusive. The Vietnamese text stayed unchanged, but intended profiles were always rated first. The 80-versus-20-point drops do not rule out order effects that differ between systems, even though the same annotators saw interleaved items. Without reversed BASE judgments, RQ2 cannot determine whether supplied profiles, English source cues, or model defaults produced the relationship-specific wording.

RQ1’s high BASE appropriateness rates left only two Gemini and four SEA-LION outputs that could improve. No overall gain was statistically detectable after Holm correction.

The explicitly post-hoc comparison supplied a complementary measurement finding: 41/44 directional choices across the original annotators’ Gem-

ini judgments favored DYAD (18/20 and 23/24; both Holm-significant), while neither SEA-LION test detected a preference. The task followed the primary analysis and reused its annotators, so this finding remains hypothesis-generating.

These findings support judging Vietnamese translations with the intended relationship visible: person reference and request wording depend on participant relations (Sidnell and Shohet, 2013; Nguyen and Ho, 2013). Our task supplies factual profiles rather than prescribed formality labels (Vakharia et al., 2023). Two translations can both fit the intended relationship yet differ in preference. Between-system differences remain descriptive because no interaction was tested.

Future work should preregister analyses, evaluate both BASE and DYAD outputs under the same fixed-text role-reversal protocol with counterbalanced evaluation order, use independent follow-up raters, and broaden the profile inventory. Comparing the reversal contrasts would help assess whether relationship sensitivity differs between translations generated with and without supplied profiles.

9 Conclusion

In this 30-request sample, Gemini’s fixed-text DYAD outputs fit 20/20 intended versus 4/20 reversed profiles; SEA-LION was inconclusive. RQ1 detected no aggregate gain. Post-hoc, 41/44 Gemini choices favored DYAD; neither SEA-LION test found a preference. Near the upper bound, binary ratings can miss pragmatic differences.

Limitations

This study uses 30 purposively selected short workplace requests in one translation direction, English to Vietnamese. The source-form strata provide descriptive balance; the sample is neither random nor representative. The controlled profiles cover only one scenario family: same-company, 30-year-old Vietnamese women who know each other professionally but are not personally close. The findings therefore do not generalize directly to other genres, age or gender combinations, degrees of familiarity, customer-facing relationships, languages, or model versions. The operational source-form strata were also not independently validated as pragmatic-strategy categories.

The experiment evaluates two LLMs and one conventional baseline; Google Translate was accessed through the unofficial `googletrans` interface rather than a frozen official API. With only two primary annotators, individual interpretations may strongly influence the aggregate results, so the findings should not be read as uniform Vietnamese norms. The combined acceptability question does not distinguish meaning errors from unnatural or grammatically unacceptable Vietnamese, and a negative pragmatic judgment does not identify which person-reference, register, or request-form feature caused it. Gemini did not expose an immutable checkpoint digest or provider-default decoding fields, and Google did not expose its backend model/version. The Ollama server build, exact invocation commands, and any attempt overrides were not recorded.

The BASE–DYAD comparison changes both relationship information and prompt length, so RQ1 does not isolate how context affects generation. The fixed-text RQ2 contrast avoids those generation differences but has a fixed-order limitation: intended judgments occurred in the main task before reversed judgments in a later module. Spacing was controlled, but reversal order was not randomized. Memory, learning, fatigue, threshold drift, demand effects, and module framing may contribute to the observed contrast. Because BASE outputs were not reversed, RQ2 does not estimate the causal effect of profile prompting. The original annotators also completed the post-hoc preference task after prior exposure to the outputs. Finally, RQ1 contains at most 30 pairs per LLM and RQ2 20; estimates are imprecise, and nonsignificance is not evidence of no effect.

Ethical Considerations

Human evaluation involved two paid adult native speakers of Vietnamese as primary annotators, referred to pseudonymously as A1 and A2, and an expert adjudicator, ADJ, who served as a research contributor. Before manuscript submission, both primary annotators provided explicit written consent to research use of their completed contributions and to reporting them in an academic paper under the pseudonymous labels A1 and A2. Their judgments are reported only as aggregate or pseudonymous rater-specific summaries, and their qualifications only collectively. The task involved ordinary language judgments and did not solicit sensitive personal information. No names, contact information, item-level judgments, free-text reasons, adjudication rationales, case or presentation identifiers, timestamps, or private audit records are released.

OpenAI tools, including Codex, were used to assist with manuscript planning and organization, code development and debugging, formatting checks, and language revision. Anthropic Claude also assisted with the camera-ready revision. All AI-assisted output was reviewed and verified by the authors, who retain full responsibility for the analyses, citations, interpretations, and claims.

References

- Ted Byrt, Janet Bishop, and John B. Carlin. 1993. [Bias, prevalence and kappa](#). *Journal of Clinical Epidemiology*, 46(5):423–429.
- C. J. Clopper and E. S. Pearson. 1934. [The use of confidence or fiducial limits illustrated in the case of the binomial](#). *Biometrika*, 26(4):404–413.
- Jacob Cohen. 1960. [A coefficient of agreement for nominal scales](#). *Educational and Psychological Measurement*, 20(1):37–46.
- Long Doan, Linh The Nguyen, Nguyen Luong Tran, Thai Hoang, and Dat Quoc Nguyen. 2021. [PhoMT: A high-quality and large-scale benchmark dataset for Vietnamese-English machine translation](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4495–4503, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Bradley Efron. 1979. [Bootstrap methods: Another look at the jackknife](#). *The Annals of Statistics*, 7(1):1–26.
- Sture Holm. 1979. [A simple sequentially rejective multiple test procedure](#). *Scandinavian Journal of Statistics*, 6(2):65–70.

- Yoko Kayano and Saku Sugawara. 2025. [Specification-aware machine translation and evaluation for purpose alignment](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 113–141. Association for Computational Linguistics.
- Ahrii Kim. 2025. [Context is ubiquitous, but rarely changes judgments: Revisiting document-level MT evaluation](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 81–97. Association for Computational Linguistics.
- Bryan Klimt and Yiming Yang. 2004. [The Enron corpus: A new dataset for email classification research](#). In *Machine Learning: ECML 2004*, pages 217–226. Springer.
- Quinn McNemar. 1947. [Note on the sampling error of the difference between correlated proportions or percentages](#). *Psychometrika*, 12(2):153–157.
- Raphael Merx, Ekaterina Vylomova, and Trevor Cohn. 2026. [Beyond “to whom it may concern”: Tailoring machine translation to audience and intent](#). *Preprint*, arXiv:2606.03259.
- Ayuna Nagato and Takuya Matsuzaki. 2025. [Character-aware English-to-Japanese translation of fictional dialogue using speaker embeddings and back-translation](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 180–190. Association for Computational Linguistics.
- Thi Thuy Minh Nguyen and Gia Anh Le Ho. 2013. [Requests and politeness in Vietnamese as a native language](#). *Pragmatics*, 23(4):685–714.
- Xing Niu, Sudha Rao, and Marine Carpuat. 2018. [Multi-task neural models for translating between styles within and across languages](#). In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1008–1021, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- Jack Sidnell and Merav Shohet. 2013. [The problem of peers in Vietnamese interaction](#). *Journal of the Royal Anthropological Institute*, 19:618–638.
- Thoai Nu-Linh Ton. 2018. [Ellipsis of terms of address and reference in casual communication events in Vietnamese](#). *Language and Linguistics*, 19(1):196–208.
- Priyesh Vakharia, Shree Vignesh S, and Pranjali Basmatkar. 2023. [Low-resource formality controlled NMT using pre-trained LM](#). In *Proceedings of the 20th International Conference on Spoken Language Translation (IWSLT 2023)*, pages 321–329, Toronto, Canada. Association for Computational Linguistics.
- Sebastian T. Vincent, Loïc Barrault, and Carolina Scarton. 2022. [Controlling extra-textual attributes about dialogue participants: A case study of English-to-Polish neural machine translation](#). In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*, pages 121–130, Ghent, Belgium. European Association for Machine Translation.
- Huy Hien Vu, Hidetaka Kamigaito, and Taro Watanabe. 2024. [Context-aware machine translation with source coreference explanation](#). *Transactions of the Association for Computational Linguistics*, 12:856–874.

A Data Availability

The research artifact is available at <https://github.com/CoolJosh0221/en-vi-dyadic-mt-artifact>. The package contains:

- the 30 selected English requests and their synthetic intended profiles;
- 199 distinct combinations of English source text, Vietnamese translation, and displayed context, covering the textual content of the 210 generated records;
- recorded generation parameters and the prompt/message structure;
- machine-readable aggregate and pseudonymous rater-specific summary tables, verification code, a manifest, and checksums.

Deduplication by source text, translation, and displayed context removes 11 duplicate entries from the 210 generated records. The package does not include the exact system/condition mapping needed to recover all 210 labeled records, or the verbatim prompt files retained privately. Its code checks inventory totals and exact-test calculations from the reported aggregate counts; it does not reconstruct those counts from item-level judgments or recompute all confidence intervals.

The primary annotators authorized research-publication use under the pseudonymous A1 and A2 labels, but public release of item-level judgments and free-text responses is outside the documented consent scope. Judgments, reasons, adjudication records, timestamps, schedules, and linkage identifiers therefore remain restricted at the item level. The intermediate source-candidate pool and record-level Enron keys, spans, and checksums remain unavailable, so candidate discovery and source-span verification cannot be replayed.