International Journal of Linguistics and Translation Studies
Volume 7, Issue 4, 2026
Homepage: http://ijlts.org/index.php/ijlts/index
DOI: https://doi.org/10.36892/ijlts.v7i4.801

How Language Models Translate Responses in Mandarin Conversation

Yang Liu
Guilin University of Technology, China
874013500@qq.com

How to cite:
Liu, Y. (2026). How Language Models Translate Responses in Mandarin Conversation. *International Journal of Linguistics and Translation Studies*, 7(4).103-118. https://doi.org/10.36892/ijlts.v7i4.801

ARTICLE HISTORY
Received: *05/09/2026*
Accepted: 12/11/2026

Keywords:
conversational translation; large language models; Mandarin Chinese; human evaluation; pragmatics

Abstract

A speaker who says "okay" and then asks a question may be checking how to carry out an agreement. An English translation can preserve both the acknowledgement and the question yet change what the speaker is asking about. This study examines how ChatGPT and DeepSeek translated responses from Mandarin conversation transcripts, with and without the preceding dialogue. Two bilingual judges consulted the Chinese conversations and found most translations consistent with the original response. The clearest changes concerned who was speaking and what a question was meant to confirm. In one exchange, a caller checked which number to dial; the English wording instead appeared to ask whether a phone call was correct. Other apparent changes were harder to establish because the Chinese and English each allowed more than one reading. For translation reviewers, the useful question is what specific wording warrants a correction. Tracing a disputed phrase to its speaker and to the action under discussion can identify a consequential change even when the English sounds plausible.

1. INTRODUCTION

When someone accepts an offer and asks how to carry it out, a translation must do more than supply a plausible equivalent of 'okay'. It must preserve who accepts, what the subsequent question concerns and how that question advances the arrangement. A response can look broadly appropriate while changing one of these details. Equally, a reviewer who knows the source may find a change in the English that another source-supported reading would allow. Both problems matter when evaluating conversational translations.

Context is an established concern in machine translation and its evaluation (Castilho & Knowles, 2025). Translation-mediated conversations make this concern particularly concrete: context can distinguish an acknowledgement from a literal rendering (Pombal, Agrawal, Zaranis, Fernandes, & Martins, 2026). The present study examines a narrower problem: what evidence warrants saying that a translated response has changed, and what that diagnosis would require a translator to revise.

Two bilingual judges compared the English responses with the Chinese conversations. Both found most translations compatible with the source. A closer look at the exceptions reveals an

important distinction: a new speaker label can unambiguously reassign an action, whereas a change in how firmly someone appears to agree can be open to interpretation.

The study asks what evidence justifies a judgement of change, what a description such as ‘acceptance followed by confirmation’ can miss, and whether preceding dialogue resolves these difficulties in the observed translations. A separate exploratory comparison examines how one local model handles the same transcripts when speakers and utterances are explicitly labelled.

2. RESPONSE RELATIONS AND EVIDENCE IN TRANSLATION EVALUATION

2.1.Interpreting the Response

Acknowledgements help organise conversation as well as convey lexical meaning. Their transitional work and the distinctions between agreement, acknowledgement and alignment caution against assigning one action to an expression in every occurrence (Beach, 1993; Wang, Tsai, Goodman, & Lin, 2010). Prior talk, accompanying contributions and subsequent treatment all bear on the interpretation.

‘Response relation’ is used here as a descriptive term for how a contribution connects to preceding talk and how its component actions connect to one another. A speaker may accept an arrangement and ask about its implementation, receive information and announce an inquiry, or evaluate a suggestion and move towards closure. These descriptions do not imply that every positive expression conveys acceptance, or that each action occupies a separate turn. A response excerpt may extend across several transcript lines and contain more than one action.

Three questions guide the analysis: who performs each action, what it concerns, and how it connects to the preceding contribution and the rest of the response. Their answers are related but not interchangeable. If a translation assigns thanks to an additional speaker, the action expressions may survive while their attribution changes. If a translation changes the object of a confirmation, acknowledgement followed by confirmation may remain a useful broad description. A reviewer must specify the level at which a difference occurs before deciding what to correct.

Action descriptions depend on context, timing and segmentation (Deppermann & Gubina, 2025). Participants’ subsequent conduct can provide evidence for a conversation-analytic claim (Raymond & Robinson, 2024). In interpreted interaction, repair sequences also make attribution of responsibility consequential (Wadensjö, 2026). Here, later Chinese turns are consulted where they clarify a response: an answer specifying contact options, for example, helps establish what a question was checking. Such evidence supports an interpretation of the Chinese sequence; it does not show how an English recipient would understand a generated translation. The analysis is text-based and sequentially informed, rather than a complete conversation analysis of every recording.

2.2.Specifying What the Evaluation Establishes

Evaluation briefs make the purpose and circumstances of human machine-translation evaluation explicit (Liu, Lo, Marshman, & Knowles, 2024). In this study, the judges could consult the full Chinese conversation; the systems received either the target alone or limited preceding context. Source knowledge may identify a problem, but may also make awkward

English easier to interpret. The analysis therefore distinguishes an output's adequacy against the full source from the information available to the system when it produced that output.

Throughout this article, 'compatible' means that a source-supported interpretation remains available in the English under bilingual review. It does not mean that this is the only possible reading, that every detail is accurate, or that a reader without Chinese would reach the same understanding. The judges assessed compatibility rather than actual communicative outcomes. No English recipient was observed using a translation to arrange a meeting, return a call or resolve a parcel inquiry.

Variability presents a further methodological difficulty in LLM translation (Rico Pérez, 2026). Collection through web interfaces did not establish fixed backend checkpoints or control sampling. Paired outputs permit inspection of particular versions with and without context, but do not isolate its causal effect. The later local experiment fixes the checkpoint and decoding settings; it changes a package of input features rather than isolating one feature's contribution.

3. MATERIALS AND PROCEDURE

3.1.Source Collection

The material comes from the DIG Mandarin Conversations (DMC) corpus (Yu, Wu, Drew, & Raymond, 2024). A local acquisition snapshot dated 7 October 2026 covered the official historical archive and listed updates through August 2024. It contained 306 transcript file records, representing 305 distinct file hashes; these are not counts of independent conversations. Supplementary Material S1 and its accompanying data identify the catalogues, original filenames, hashes and transcript locations.

After excluding 23 files from an inherited development set, four written forms were searched in the remaining 282 files: 好的 (hao de), 行啊 (xing a), 那行 (na xing) and 可以啊 (keyi a). Literal string searches were supplemented by a Chinese-character projection removing intervening non-Chinese characters. This retrieved locations for contextual inspection, not an exhaustive collection of Mandarin acceptances or acknowledgements. Explicit forms can help locate pragmatic phenomena, but do not substitute for contextual analysis (Nagy C., Németh T., & Németh, 2025).

The search matched 127 files. Seven subsequently used in methodological development were also set aside, leaving 120 files for screening. Preliminary AI-assisted screening sought a discernible positive or receipt component together with another action-bearing contribution in the same response; thanks and closing contributions were eligible. Of 245 matching locations, 81 were provisionally identified as compound-response anchors. The remainder comprised lexical or predicative uses, uncertain cases, repetitions or single actions, and metadata. Merging anchors within the same response produced 78 candidates from 58 files. These were screening decisions, not independently validated action categories; the complete accounting is retained in S1.

One candidate contained alternating speakers inside the proposed target envelope. It was held before generation because the fixed request called for one target response, and removing the intervening speech would alter the source. The translation cohort therefore contained 77 targets from 57 files, or 56 conservative event groups because two files may derive from the same

cake-delivery event. The unit was the selected response excerpt, not necessarily a complete acoustic turn or a uniform instance of proposal acceptance.

Later source review identified an incomplete tail and two targets from a conversation containing a self-identified robot. These remain in the primary collection and are removed in a sensitivity analysis. The cohort was not wholly unseen during development: 32 of the 120 screened files had appeared in earlier source or AI work, six historical reviews could not be confidently matched to files, and seven final targets had been used in human calibration. These exposures are documented. The study was not preregistered, and the interpretive distinctions used in the close analysis were refined after collection.

3.2.Translation Requests

Each target had two input conditions. C0 supplied the target with an empty context field. C1 added up to eight preceding speaker blocks, or fewer if less preceding material was available. A block was introduced by a speaker label and could span several printed lines; it was not a fixed number of sentences or seconds. Subsequent talk was excluded. Overlap brackets, pauses, cut-offs and embedded romanised fragments were retained from the spoken transcript layer; separate pinyin and gloss layers were removed.

The instruction, identical across conditions, requested translation of the target into English, output of that translation only, and neither translation of the context nor explanation. Context and target occupied separate tagged fields. S1 provides the original Chinese wording and all 154 packets. The prompt's term 'turn' did not determine whether every excerpt constituted a complete turn analytically. No preferred action interpretation or reference translation was supplied.

The packets followed one fixed random order in both products, with a new conversation for each request. The seed, 2026101002, controlled packet order, not model sampling. October 2026 collection used ChatGPT with the interface set to Medium and DeepSeek with DeepThink enabled and Search disabled. 'ChatGPT' and 'DeepSeek' below refer to these observed web-product configurations, not independently verified model checkpoints. The settings are not treated as equivalent reasoning budgets.

Of 308 planned first attempts, 307 yielded captured answers. One DeepSeek C0 attempt failed; a later retry was retained separately. No preferred output was selected from repeated successful generations. Interface access difficulties required the author to copy and paste four ChatGPT answers manually. One DeepSeek submission contained an extra trailing character. The primary accounting retains these five discrepancies; a sensitivity analysis excludes affected targets to preserve clean, complete four-output sets. Exact submitted prompts and fine-grained times are not independently recoverable for every ChatGPT record.

Temperature, sampling seed, hidden routing and server-side changes were uncontrolled. Fresh conversations reduce visible carry-over but do not rule out account-level influences or prior service exposure to the corpus. These limitations concern reproducibility of generation, not the availability of the captured texts for scrutiny. The study compares particular first answers rather than distributions from repeated sampling.

3.3.Human Review

Two non-author bilingual judges, A and B, reviewed the Chinese sources and then the translations. Both were Master of Translation and Interpreting (MTI) students with Test for English Majors Band 8 qualifications. The author supplied none of the formal ratings. The source document contained all 78 candidates, including the held candidate, with their full available conversations. Questions asked what the response addressed, what its parts did, how those parts connected, and whether its interpretation or boundary remained uncertain.

The English document contained 289 items. Exact English duplicates were merged only within a source target: 307 first-output records yielded 288 distinct target–English items, and the retry added one item. Identical sentences from different targets were not merged. Each item linked to its Chinese source and asked separately about content and response relations. Product and condition labels were concealed; both judges received the same fixed shuffled order. Concealment could not prevent inference from style or comparison between versions of the same target.

Content judgements distinguished no substantive difference, a difference, uncertainty and inability to evaluate. The question covered participants, actions, polarity, time, quantity and decision scope. The relation question asked whether a source-supported interpretation remained compatible with the English or whether particular wording supported a change; uncertainty and inability to evaluate were also available. These dimensions overlap. Reassigning an action to another person can affect both content and the response as a contribution to the interaction. The questions do not constitute independent semantic and pragmatic scales.

Both documents were delivered together. The author confirmed that both judges completed source review first, neither consulted the other's answers or discussed the questions during completion, and neither used AI or translation tools, including for rewriting. These are procedural confirmations, not independently monitored records of completion. The author's earlier development calibration and selective listening, including the parcel case below, are distinct from A's and B's formal answers. Calibration notices in the materials record earlier project use, not either judge's personal exposure. The two judges did not independently review audio for every candidate.

Answers were retained without consensus adjudication. Their explanations have an important limitation: across all 289 items, A repeated one general evidence sentence in 271 answers; in 259 answers, B used a fixed introduction followed by the exact wording of the corresponding Chinese source-function answer. B's reasons thus varied across sources but often failed to identify wording specific to an English version. These patterns reduce the item-specific evidence for compatibility. They do not establish the process by which a label was reached or disprove the reported independence of completion. S1 documents this audit and preserves the original instructions and answers.

3.4.Analysis

Overall counts use the 288 distinct first-answer items. Product-by-condition counts use output records, because identical outputs from multiple conditions can map to one judged item. Context comparisons are paired within target and product. The failed first attempt remains

missing; the retry is not substituted for it. No significance test treats these clustered records as independent conversations.

All 34 distinct first-answer items receiving a content difference, content uncertainty or relation change from either judge were checked against the source. The close analysis also includes compatible examples and full evidence chains for 15 targets. This review was AI-assisted, not a third human judgement. Where the analysis qualifies an interpretation, the original label remains in the counts. The compatible examples were selected for exposition rather than sampled to estimate overlooked changes.

Agreement is reported overall and for the infrequent change category. Positive agreement is twice the number jointly marked changed divided by the sum of each judge's change count. Sensitivity analyses remove incomplete product sets, capture discrepancies, previous calibration targets and named source caveats. They describe dependence on inclusion decisions; they do not establish representativeness or independence.

4. HOW THE JUDGES ASSESSED THE TRANSLATIONS

Both judges found the response relation compatible in 274 of 288 first-answer items (95.1 per cent), jointly marked seven as changed (2.4 per cent) and disagreed on seven (2.4 per cent). Percentages do not sum to 100 because of rounding. All disagreements were compatible judgements from A and change judgements from B. Table 1 reports individual distributions, including content judgements.

Table 1. Individual Judgements on 288 Distinct First-Answer Items

Judgement	Judge A	Judge B
Content: no substantive difference	270	258
Content: difference	10	27
Content: uncertain	8	3
Relation: compatible	281	274
Relation: changed	7	14

Note. Both judges are non-authors. Each dimension totals 288 per judge. No relation-uncertain or unevaluable responses occurred in this set. The retry is excluded.

Relation agreement was 281/288 (97.6 per cent); content agreement was 267/288 (92.7 per cent). Overall agreement is dominated by compatible items. Positive agreement on change was $2 \times 7/(7 + 14)$, or 66.7 per cent. The seven jointly marked items represent six source targets; the 14 marked by either judge represent nine. These are judgements on this material, not accuracy estimates for conversational translation.

Every relation-change mark was accompanied by a content-difference mark from the same judge. The overlapping questions cannot establish the absence of purely pragmatic changes.

They do, however, provide no positive evidence that propositional content was intact while response relations alone were damaged.

Table 2. Relation-Change Marks by Product and Input

Product	Input	Outputs	Both judges	Either judge
ChatGPT	C0	77	1	2
ChatGPT	C1	77	0	2
DeepSeek	C0	76	4	7
DeepSeek	C1	77	2	3

Note. C0: target only; C1: target and preceding dialogue. The last two columns count outputs marked changed. Counts concern output records, not independent judgements; exact within-target duplicates reuse one judgement. The failed DeepSeek C0 attempt and its retry are excluded.

Among 77 complete ChatGPT pairs, one target was jointly marked changed in C0 and compatible in C1; 76 had no joint change in either condition. Among 76 complete DeepSeek pairs, three moved from a joint change in C0 to compatibility in C1, one acquired a joint change mark in C1, one remained jointly changed and 71 had no joint change. The newly joint C1 mark had already been a B-only mark in C0, so the shift in agreement should not be reported as an entirely new concern.

The sensitivity results in S1 retain the predominance of joint compatibility. Applying all restrictions leaves 61 targets and 229 items: 219 jointly compatible, four jointly changed and six disagreements. Complete-set restrictions remove a target's otherwise usable outputs along with the affected output. The reduction makes dependence on inclusion decisions visible but leaves too few joint changes to support a product ranking.

5. WHAT CHANGED IN TRANSLATION

The examples include compatible translations and all six targets with joint change marks. Chinese quotations retain transcript notation. Glosses and free translations are explanatory aids, not scoring references. Generated English is reproduced without correcting wording or punctuation; surplus horizontal spaces are collapsed. Square brackets mark source overlap, colons lengthening, hyphens cut-offs, parenthesised numbers pauses, (.) a brief pause, and degree signs quiet speech. An equals sign marks latching, an upward arrow a pitch rise, angle brackets tempo changes, and double parentheses nonlexical events. These marks in generated English do not establish English prosody. Item identifiers distinguish rated texts.

5.1.Compatible Responses

In (1), Wan asks where the participants should eat later. Zhai leaves the location open and asks for Wan's preference. Both judges interpreted the response this way. The dormitory is proposed only subsequently: treating the target as acceptance of that location would import a later decision.

(1) Leaving the choice of location open (SQ007-01)

Chinese target

宅: [(嗯.)

(0.8)

呃:::都行啊. ((咳嗽)) 你想在哪儿吃啊,

Zhai mm uh all fine particle cough you want at where eat particle

Zhai: (Mm.) Uh, anywhere is fine. ((coughs)) Where do you want to eat?

ChatGPT C0 (E157)

Mm... Uh, anywhere's fine. Where do you wanna eat?

Both judges accepted all four translations. The wording varies, but Zhai leaves the location open and asks where Wan wants to eat.

In (2), a caller receives contact information, undertakes to contact the delivery person immediately and adds thanks. The undertaking, timing and gratitude are overt in the target, although preceding talk identifies 'him'.

(2) Receiving information and undertaking to act (SQ043-01)

Chinese target

A: ↑噢:::好的. 好的. 我 现 在 联 系他.谢 谢.

A oh okay okay I now contact him thank you

A: Oh, okay. Okay. I will contact him right now. Thank you.

ChatGPT C0/C1 (E214)

Oh, okay. Okay. I'll contact him right now. Thank you.

The identical ChatGPT outputs counted as one item. Both judges also accepted the DeepSeek versions: in each, the caller undertakes to make contact.

5.2.Reassigning Speakers and Roles

In (3), Qiqi has offered to obtain students' telephone numbers and send them through WeChat. Xiaoqian accepts and thanks Qiqi. The marked Chinese response belongs entirely to Xiaoqian; both source judgements placed the thanks after acceptance of the promised help.

(3) Acceptance and thanks assigned to different speakers (SQ003-01)

Chinese target

小浅: 嗯， 可以可以.好的.

[行:， 谢谢你.

Xiaoqian mm can can okay fine thank you

Xiaoqian: Mm, sure, sure. Okay. Alright, thank you.

DeepSeek C0 (E211)

Xiaoqian: Mm, sure, sure. Okay. Xing: Thank you.

DeepSeek C1 (E153)

Xiaoqian: Mm, okay, okay. Alright. [Okay, thank you.

In C0, ‘Xing:’ gives the thanks to a second speaker. In the Chinese transcript, 行 belongs to Xiaoqian’s utterance. Both judges marked the change. With preceding dialogue, DeepSeek rendered 行 as ‘Okay’ and kept the thanks with Xiaoqian.

A related split remains with context in (4). The participants discuss posted swimming equipment. Ping says that its recipient will be able to swim instead of continuing to run. Qiang agrees about the benefit for his knees and then says he was only asking, moving towards closure of the inquiry.

(4) An invented speaker at the move towards closure (SQ028-02)

Chinese target

强: 对. 不然太伤膝盖了. °我的天啊. °

那:那行. 我就只是问一下, 哈.

Qiang right otherwise too harm knees particle my heavens particle then then fine I just only ask a bit ha

Qiang: Right. Otherwise it is too hard on the knees. Oh my goodness. Well, alright. I was just asking, ha.

DeepSeek C0 (E269)

Qiang: Right. Otherwise it’s too hard on the knees. (Oh my god.)
Na: Alright then. I was just asking, ha.

DeepSeek C1 (E215)

Qiang: Right. Otherwise it’s too hard on the knees. °Oh my god.°
Na: Alright then. I was just asking, ha.

Both DeepSeek translations turn the lengthened 那 (na) into a speaker label, ‘Na:’. Qiang’s remark that he was just asking now appears to come from someone else. Here, providing the preceding dialogue did not prevent the split.

An institutional role can also change without an additional speaker. In (5), a service representative announces a transfer. The source abbreviation 客1 is difficult to interpret in isolation, whereas the wider conversation identifies an agent performing the transfer.

(5) Reassigning the role of the person making a transfer (SQ024-01)

Chinese target

客1: 哎(.)那行.哎.转[接中.

Agent one oh then fine oh transferring

Agent 1: Oh, alright then. Oh, transferring now.

DeepSeek C0 (E165)

Customer 1: Oh (.) alright then. Oh. Transfer[ring.

DeepSeek C1 (E060)

Agent 1: Oh (.) okay then. Oh. Transferring [now.

The transfer itself survives in both translations, but ‘Customer 1’ assigns it to the wrong party. The C1 label, ‘Agent 1’, identifies the person responsible.

Both ChatGPT outputs omitted the role label and were judged compatible. Because one was copied manually, this target drops out of the sensitivity analysis that requires complete sets without capture discrepancies.

5.3.Changing the Object of Confirmation

In (6), Lin has agreed to remind her roommate to contact a teacher. The teacher says the student can return the call if convenient. Lin accepts and checks the contact detail. The teacher’s subsequent answer confirms it and offers a mobile number and QQ as alternatives. This continuation supports interpreting dianhua as a way to reach the teacher, most naturally the current number.

(6) Confirming a number or evaluating a call (SQ018-02)

Preceding turn

师: 啊:.好的好的.就是方便的话让她给我回个电话也可以.=

Teacher ah okay okay just convenient if let her to me return a call also can

Teacher: Ah, okay, okay. If it is convenient, having her call me back would also be fine.

Chinese target

林: =啊::.好的好的.就这个电[话是吗.

Lin ah okay okay just this telephone is question particle

Lin: Ah, okay, okay. This telephone number, right?

DeepSeek C1 (E037)

Lin: =Ah::. Okay, okay. Is this phone call [right?

DeepSeek C0 (E094)

Lin: =Ah::. Okay okay. It’s just this pho[ne, right.

ChatGPT C1 (E257)

Oh, okay, okay. This number, right?

The Chinese question checks where to call back. DeepSeek C1’s ‘Is this phone call [right?’ instead appears to ask whether the call itself is correct. Both judges marked a change even though the English still contains an acknowledgement followed by a question.

The judges split on C0: ‘this phone’ could stand for the number. Both accepted the ChatGPT versions, which specify ‘number’. The teacher’s subsequent answer helps establish Lin’s meaning; only the preceding request for a callback was available to the model in C1.

The parcel inquiry in (7) involves another confirmation. After difficulty identifying a missing parcel, its recipient confirms that the item is a thermos. The postal worker announces an inquiry to the parcel station and repeats the item description. The recipient then agrees before asking whether there is a problem. The written sequence, supplemented by the author’s development listening, supports treating the final component as checking parcel information relevant to the inquiry. It does not settle every intonational detail.

(7) Parcel identification rendered as buyer identification (SQ058-01)

Chinese target

邮: °哦:::. °那行我问一下菜鸟驿站.<是si-买了个暖瓶啊.

Postal worker oh then fine I ask a bit Cainiao parcel station is si bought a thermos particle

Postal worker: Oh, alright, I will ask the Cainiao parcel station. It is—[you] bought a thermos, huh.

Following turn

凡: 呃对.是有是有什么问题吗, °hehe°

Fan uh right is have is have what problem question particle laugh

Fan: Uh, yes. Is there, is there some problem? ((laughs))

DeepSeek C0 (E173)

You: °Oh:::. °All right then, I'll ask the Cainiao Station.<It was si- who bought a thermos.

DeepSeek C1 (E004)

邮: °Oh:::. °Alright, I'll ask Cainiao Station. <Did [you] buy a thermos?

The postal worker has just said that he will ask the parcel station about the item. ‘It was si- who bought a thermos’ turns a check on the parcel’s contents into a statement about a buyer. Both judges marked a change. The cleft construction, rather than the ambiguous label ‘You:’, is the evidence for that reading.

Both judges accepted C1. ‘Did [you] buy a thermos?’ can still serve to identify the item the postal worker is looking for.

5.4.Qualifying a Joint Change Mark

The last jointly marked target concerns where to go. Jiong responds to Ke’s account of the journey, comments that visiting only the mall seems uninteresting and conditionally proposes Shilaoren. The expression 那听你这个样 is colloquial and incomplete, but the final proposal explicitly makes the trip conditional on the inconvenience involved.

(8) A possible strengthening of uptake (SQ015-01)

Chinese target

炯: 噢:那行.=那听你这个样=(>然后<)反正我是觉得

出来只看这商场好像也没什么意思.=那你如果不

嫌麻烦那不然就去石老人吧.

Jiong oh then fine then hear you this way then anyway I think come out only look this mall seems also not much meaning then you if not mind trouble then otherwise just go Shilaoren particle

Jiong: Oh, alright. From what you are saying—(then) anyway, I feel that coming out just to look around this mall is not very interesting. If you do not mind the trouble, why not go to Shilaoren?

ChatGPT C0 (E196)

Oh, okay then. We'll go with what you suggest. Anyway, I don't think there's much point in going out just to look around the mall. So if it's not too much trouble for you, why don't we go to Shilaoren instead?

ChatGPT C1 (E138)

Oh, okay then. Well, since you put it that way, I mean, I don't think there's much point in coming all this way just to look around this mall. So if it's not too much trouble for you, why don't we go to Shilaoren instead?

Both judges marked C0 as changed. 'We'll go with what you suggest' does sound more definite, but Ke had already mentioned Shilaoren, and the English retains Jiong's conditional proposal. The evidence supports a possible strengthening of commitment at one point, rather than a reversal of the whole response.

Both judges accepted C1's 'since you put it that way'. The joint C0 mark remains in the totals.

5.5.Disagreement and Competing Readings

The seven disagreements concern four targets. Three involve 想太多了 rendered as 'You're overthinking it' in an assessment of a dormitory-moving plan followed by call closure. The Chinese can treat the plan as wishful or unrealistic rather than diagnose excessive thought. B marked the versions changed and A accepted them. A critical stance and ensuing closure remain available, so these are not three established failures to preserve the entire response sequence.

Two disagreements concern 我过来了, translated as 'I'm on my way' or 'I'm coming over.' Both source answers favoured the speaker having come over, but earlier talk places him in transit. The supplied transcript fixes neither elapsed time nor arrival at the meeting point. A plan to wait by the road survives in both versions. A definitive temporal-error claim would require stronger source evidence.

The remaining disagreements concern 'this phone' in (6) and ChatGPT C1's 'I'll take it out when it arrives' for 我到时候把它拿出来. In the latter, preceding talk concerns a parcel's expected arrival and contents. 'When it arrives' is more specific than 'when the time comes,'

but plausibly draws on the supplied context. B's mark is retained without treating it as a secure instance of invented timing.

These cases distinguish ambiguity in English from indeterminacy in the Chinese evidence. A translated expression may allow competing readings, or the source may not establish the exact temporal or stance interpretation required to diagnose an error. Consensus labels would conceal that difference. Preserving the answers alongside the analysis makes it possible to see both the concern and its evidential limit.

6. WHEN DOES A TRANSLATION NEED CORRECTION?

A useful correction can be quite small. Removing 'Xing:' restores the connection between Xiaoqian's acceptance and her thanks. Replacing 'phone call' with 'number' restores the purpose of Lin's question. Neither correction requires changing the acknowledgement. Table 3 distinguishes these observable changes from readings on which the source leaves room for doubt.

Table 3. Translation Problems and the Evidence for Correction

What to check	Evidence in the English	What the source supports
Speaker or role	"Xing:" before the thanks (3); "Na:" before closure (4); "Customer 1" during a transfer (5).	One person accepts and thanks; one person closes the inquiry; the service agent makes the transfer.
Object of confirmation	"Is this phone call [right?" (6); "who bought a thermos" (7).	The caller checks a number. The postal worker checks which item to ask the parcel station about.
Strength of commitment	"We'll go with what you suggest" (8). Both judges marked a change.	The wording suggests firmer agreement, but the conditional proposal still leaves the destination open.
Timing or stance	"I'm on my way" and "You're overthinking it" divided the judges (Section 5.5).	The exchange does not settle the precise time of arrival; criticism of the plan remains a possible reading.

Every judge who marked a change in the response relation also marked a content difference in that item. Counting these as two independent errors would misdescribe the evidence: the choice of speaker or referent is precisely what changes the relation.

For a reviewer, the first step is to locate the response that was meant to be translated. If the output covers that response, compare the speaker and the object of each action. Before proposing an edit, identify the words that rule out the source reading. If the available talk supports both readings, as in the discussion of Shilaoren, a categorical correction is difficult to defend. The study did not test whether these checks improve professional editing decisions or save time.

The judges could consult the Chinese, and their written explanations often repeated the same rationale across items. Joint compatibility therefore indicates that the judges could reconcile the translation with the source; it does not establish how often an English reader would do so.

Preceding dialogue accompanied the recovery of Xiaoqian's speaker label, but also the invented 'Na:' and the question about a 'phone call'. These pairs cannot separate the effect of context from variation between generations. They motivated a narrower follow-up: what happens if the input marks the speakers and the target explicitly?

7. AN EXPLORATORY COMPARISON OF INPUT REPRESENTATIONS

The follow-up used all 77 targets with their saved C1 preceding windows. One condition retained the plain transcript in tagged fields. The other used JSON speaker and utterance fields plus a generic guide to notation, including lengthening, overlap and incomplete romanised fragments. Both requested only the target speaker's English utterance, without labels, explanation or background translation. This output instruction differed from the earlier web-product request. The comparison therefore concerns the two new input packages, not performance against ChatGPT or DeepSeek.

The local comparison used Qwen3-4B-Instruct-2507 in four-bit form with greedy decoding and one attempt per packet. All 154 generations finished without empty answers or truncation at the token limit. Supplement S2 records the checkpoint revision, settings, packet order and two preliminary checks on synthetic inputs. The notation guide was revised after the first check; neither check validated the guide on conversational data.

AI-assisted inspection of every pair identified substantial substitution of preceding talk for the target in 11 structured outputs and no raw outputs. This meant that target actions were omitted and identifiable earlier talk was translated instead; lesser context-based additions were recorded separately. Seven structured outputs, and no raw outputs, retained Chinese characters. These are narrow observations, not overall correctness scores. No new human ratings were collected. Excluding the two source-role-uncertain targets leaves ten substitutions among 75 structured outputs and none among 75 raw outputs.

For example, SQ010-01 ends a food-order adjustment with 顾: 嗯:好的,[好的.谢谢. The raw output is 'Um, okay, thanks.' The structured output is 'Um, okay, the same price, besides the shiitake and greens, are there any others?' This translates an earlier menu question instead of preserving the target's thanks. The intended target remains correctly delimited in the saved structured packet. Correct preparation and successful execution therefore did not ensure that the generated text respected the target boundary.

Explicit labels and the notation guide were introduced together, so the substitutions cannot be attributed to either feature alone. This comparison offers no evidence for recommending the combined format.

8. CONCLUSION

Most translations in this collection allowed both bilingual judges to recover a reading consistent with the Chinese. The clearest exceptions changed who spoke or what a question confirmed. A reviewer can look for those changes even when individual words such as 'okay' are well

translated. Where the original exchange leaves timing or commitment open, the review should preserve that uncertainty.

The findings concern selected responses found through four expressions, not Mandarin conversation in general. The judges knew the source, and jointly compatible items were not randomly audited for missed changes. The web outputs cannot establish a model ranking or a causal benefit of context. The local follow-up, without independent human evaluation, illustrates one further check: verify that the output translates the requested response before judging its wording.

DECLARATIONS

Ethics and consent

The author's institution determined that ethical approval was not required. Both non-author judges consented to research use and anonymous reporting or quotation of their answers. Existing DMC materials are attributed to their source; no new telephone conversations were recorded.

Funding and competing interests

This research received no funding. The author declares no competing interests.

Data availability

Supplementary Materials S1 and S2 provide source catalogues and locators, screening records, issued instructions, captured outputs, both judges' verbatim answers, and separately labelled analyst observations. S1 includes a script recomputing the human-rating results. S2 includes the frozen local design, prepared and tokenised requests, all outputs, development checks, model-file hashes and an offline verification script. Intended packets remain distinct from evidence of actual web submission. The review package excludes account-linked conversation URLs and personal filesystem paths and has not been publicly deposited. Original archive files and recordings remain with the DMC provider; no unrestricted redistribution right is claimed.

Generative AI use

OpenAI ChatGPT/Codex assisted with analysis code, qualitative interpretation and language revision. ChatGPT and DeepSeek generated the web-product translations; a local Qwen checkpoint generated the follow-up outputs. These uses are distinct from the two judges' answers, reported by the judges as completed without AI or translation tools. AI-assisted analysis was not counted as another human judgement and did not replace original labels. The human author retains authorship and responsibility for the submitted work.

REFERENCES

Beach, W. A. (1993). Transitional regularities for "casual" "okay" usages. *Journal of Pragmatics, 19*(4), 325–352. https://doi.org/10.1016/0378-2166(93)90092-4

Castilho, S., & Knowles, R. (2025). A survey of context in neural machine translation and its evaluation. *Natural Language Processing, 31*(4), 986–1016. https://doi.org/10.1017/nlp.2024.7

Deppermann, A., & Gubina, A. (2025). Coding actions in social interaction: Potentials and problems. *Research on Language and Social Interaction, 58*(3), 258–280. https://doi.org/10.1080/08351813.2025.2528494

Liu, T., Lo, C.-k., Marshman, E., & Knowles, R. (2024). Evaluation briefs: Drawing on translation studies for human evaluation of MT. In R. Knowles, A. Eriguchi, & S. Goel (Eds.), *Proceedings of the 16th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)* (pp. 190–208). Chicago, IL: Association for Machine Translation in the Americas. Retrieved from https://aclanthology.org/2024.amta-research.17/

Nagy C., K., Németh T., E., & Németh, Z. (2025). How to investigate implicit pragmatic phenomena in corpora. *Journal of Pragmatics, 240*, 79–90. https://doi.org/10.1016/j.pragma.2025.02.007

Pombal, J., Agrawal, S., Zaranis, E., Fernandes, P., & Martins, A. F. T. (2026). A context-aware framework for translation-mediated conversations. *Transactions of the Association for Computational Linguistics, 14*, 562–587. https://doi.org/10.1162/tacl.a.639

Raymond, C. W., & Robinson, J. D. (2024). Evidencing conversation-analytic claims: How participants orient to social action. In J. D. Robinson, R. Clift, K. H. Kendrick, & C. W. Raymond (Eds.), *The Cambridge handbook of methods in conversation analysis* (pp. 315–355). Cambridge, England: Cambridge University Press. https://doi.org/10.1017/9781108936583.013

Rico Pérez, C. (2026). Taming the stochastic parrot: A translation studies approach to variability in LLM-based machine translation. *Digital Translation*. Advance online publication. https://doi.org/10.1075/dt.25027.ric

Wadensjö, C. (2026). Some thoughts on repair in interpreter-mediated asylum interviews. *inTRAlinea*, Special Issue: Interpreting in interaction, interaction in interpreting. Retrieved from https://www.intralinea.org/specials/article/2702

Wang, Y.-F., Tsai, P.-H., Goodman, D., & Lin, M.-Y. (2010). Agreement, acknowledgment, and alignment: The discourse-pragmatic functions of hao and dui in Taiwan Mandarin conversation. *Discourse Studies, 12*(2), 241–267. https://doi.org/10.1177/1461445609346922

Yu, G., Wu, Y., Drew, P., & Raymond, C. W. (2024). The DIG Mandarin Conversations (DMC) Corpus: Mundane phone calls in Mandarin Chinese as resources for research and teaching. *Chinese Language and Discourse, 15*(1), 105–141. https://doi.org/10.1075/cld.23001.guo