Article

Check for updates

Journal of International Students
Volume 16, Issue 13 (2026), pp. 273-304
ISSN: 2162-3104 (Print), 2166-3750 (Online)
jistudents.org
https://doi.org/10.32674/2q8f1w40

Exploring the Role of Large Language Models during the Culture Shock Phase: A Study of International Students in Japan

ChunPi Hsieh[1*], Yoichi Ochiai[2, 3], Ichiro Matsuda[1], Tatsuki Fushimi[2, 3], and Jingjing Li[2*]

[1]*Graduate School of Comprehensive Human Sciences, University of Tsukuba;*
[2]*Institute of Library, Information and Media Science, University of Tsukuba;*
[3]*Tsukuba Institute for Advanced Research (TIAR), University of Tsukuba*

[*]*Corresponding authors: ChunPi Hsieh, Email: jocelyn85419@gmail.com, ORCID ID: 0009-0004-0579-6797 and Jingjing Li, Email: li@digitalnature.slis.tsukuba.ac.jp, ORCID ID: 0000-0002-6524-3105*

ABSTRACT: *International students face cultural challenges during culture shock, leading to distress and social isolation. University support lacks personalization, leaving students to rely on informal tools such as translation apps and social media that perform poorly in complex situations. Large Language Models offer multilingual support, yet their role in international students' cultural adaptation remains underexplored. This study examines how LLMs support culture-related everyday tasks in Japan, where implicit institutional norms intensify difficulties. We conducted a two-phase study: a pre-survey (N = 103) identified barriers and informed four scenarios (hair appointments, clinic visits, bank account opening, and garbage classification), followed by a user study (N = 40) comparing GPT-4o, Claude, Gemini, and Google Translate. Results show that breakdowns stemmed less from vocabulary than from uncertainty about institutional expectations and appropriate action. Participants used LLMs not only for translation but also for reassurance and sensemaking.*

Keywords: Large Language Models; Cultural Shock Phase; Cultural Adaptation; International Student; Japan; Life Support

INTRODUCTION

The increasing international mobility of students reflects a growing interest in academic and personal development abroad. In Japan, the international student population reached 336,708 by the end of 2024, and the government has set a target to increase this number to 400,000 by 2033 (Japan Student Services Organization, 2024; Nikkei, 2023). While international education offers numerous benefits, students often encounter difficulties adapting to unfamiliar cultural environments. Cultural adaptation, as Sumer (2009) defines, involves maintaining one's cultural identity while adjusting to a new one. To understand this process, Oberg's (1960) Culture Adjustment Stages theory outlines four phases: honeymoon, culture shock, recovery, and adjustment. After the initial excitement of the honeymoon phase, many students enter the "culture shock" phase, a critical period marked by emotional strain and the loss of familiar social signs. During this phase, students often experience uncertainty, anxiety, and a loss of confidence in navigating everyday situations (Jayasinghe & Rathnayake, 2022). They require not only information but also support that is immediate, personalized and capable of providing reassurance in unfamiliar contexts. These difficulties extend beyond adaptation itself and can affect psychological well-being, academic engagement, and even the willingness to remain in the host country (Xia, 2009). In the context of rapid AI development, large language models (LLMs) introduce a new possibility for addressing these needs by offering continuous, personalized, and interactive support. Many international students have already begun using such tools to cope with everyday challenges. However, from a research perspective, we still lack a clear understanding of how international students interact with AI in real-world situations, particularly during the culture shock phase. Understanding this interaction is therefore critical in the AI era, as it provides a foundational basis for supporting international students more effectively. It also offers important insights for researchers, educators, and practitioners seeking to design future support systems that align with students' real needs in everyday life.

Within Japan, this culture shock phase is intensified by the country's complex bureaucratic procedures and highly context-dependent communication norms (J. S. Lee, 2017; Murphy-Shigematsu, 2002). Rather than relying on explicit guidelines, both institutional procedures and everyday interactions often depend on unspoken rules and implicit expectations, a dynamic related to the cultural practice of kuuki wo yomu ('reading the air') (Nippoda, 2012; Wu & Ishii, 2025). Crucially, these expectations shape a range of newcomer-facing situations, including not only formal institutional contexts such as banking, healthcare, and public administration but also everyday interactional tasks such as making phone calls or communicating with service staff. For newcomers, these challenges are less about vocabulary and more about interpreting what action is required in a given situation (Murphy-Shigematsu, 2002; Ward & Kennedy, 1999). Even when linguistic meaning is understood, uncertainty about expectations, such as how to respond appropriately, what information is required, or how to proceed through multi-step interactions, can lead to hesitation, errors,

or breakdowns in communication. If left unaddressed, repeated difficulty in navigating such situations may contribute to stress, withdrawal, and reduced academic engagement (Jayasinghe & Rathnayake, 2022).

Existing support systems in Japanese universities, such as tutor programs and counseling centers, provide essential aid but are often time-bound and administratively focused (Asian Students Cultural Association [ABK] & Benesse Corporation, n.d.; Study in Japan, n.d.; Tran et al., 2022; University of Tsukuba, 2024a; Yonezawa, 2020). Research indicates that cultural adjustment challenges are most acute during unstructured real-life moments (Ward et al., 2001), such as navigating clinics or hair salons, when formal offices are closed and no immediate assistance is available. As a result, students frequently adopt technology-mediated coping strategies, relying on informal tools such as translation apps, which often fail to convey the nuance and politeness levels required by Japanese institutional logic.

The rise of LLMs offers a potential solution by providing real-time, multilingual, and context-aware assistance through natural language interaction. International students increasingly turn to AI for daily life use (Kamalov et al., 2023), leveraging these tools for key adaptation needs such as language practice, cultural scenario rehearsal, emotional reassurance, and information seeking. However, the current literature remains predominantly focused on academic writing or clinical mental health (Colace et al., 2018; Mekni, 2021; Merikko & Silvola, 2024; Park & Ahn, 2024; Wang et al., 2024). Consequently, a significant gap remains regarding how LLMs support everyday sociocultural adaptation in real-world interactional contexts, particularly during the early stages of cultural uncertainty faced by short-term international students (Masgoret, 2006; Sydoruk, 2024).

To address this gap, this study examines how existing LLM systems support international students in Japan during the cultural shock phase through a two-phase mixed-methods design. Phase 1 ($N = 103$) identified key adaptation challenges, which informed the design of four real-world scenarios for Phase 2. Phase 2 involved 40 participants who had arrived within the past six months, a timeframe identified by Ward et al. 1998 as the peak of sociocultural difficulty. Participants were randomly assigned to use GPT-4o, Claude, Gemini, or Google Translate to complete culturally specific tasks. Data sources included screen recordings, post-experiment surveys, staff evaluations, and semi-structured interviews, enabling analysis of both user needs and system performance under culturally unfamiliar conditions. This study addresses two research questions:

- RQ1: What types of support do international students seek from LLM systems during the culture shock phase?
- RQ2: Can existing LLM tools effectively support international students in accomplishing daily life tasks during the culture shock phase?

By integrating subjective perceptions with objective task outcomes, this study extends research on international student adaptation by showing that LLM use during the culture shock phase is valued less for task completion and more for helping students navigate cultural uncertainty. In doing so, it provides new insight into how AI-mediated support shapes international students' early adaptation experiences.

LITERATURE REVIEW

Sociocultural Adaptation in The Culture Shock Phase

This study is grounded in Oberg's (1960) foundational framework of culture shock, particularly the crisis phase characterized by anxiety resulting from the loss of familiar social cues. Expanding on this, Searle and Ward (1990) and Ward and Kennedy (1999) distinguished the acculturation process into two distinct domains: psychological adjustment, which refers to emotional well-being and satisfaction, and sociocultural adaptation, which relates to the behavioral competence required to execute daily tasks effectively.

The present study treats sociocultural adaptation as the primary construct, with a specific focus on the development of functional and interactional competence required to perform everyday institutional tasks in unfamiliar environments. Within this framework, culture shock is understood as the contextual condition that gives rise to these adaptation challenges, rather than the main analytical lens. Research indicates that sociocultural learning demands are most intensive immediately after arrival, with behavioral difficulties typically peaking within the first six months (Ward et al., 1998). Building on this finding, the present study operationalizes this theoretical insight through its participant selection strategy by focusing on international students within this early period. During this time, behavioral difficulties are most evident in performing concrete, institutionally structured daily tasks.

In the context of Japan, acquiring this behavioral competence is uniquely challenging due to the high-context nature of social interaction. Wu and Ishii (2025) highlighted that the Japanese institutional landscape is characterized by significant "information asymmetry" and a lack of explicit criteria, forcing students to rely on implicit cues to navigate uncertainty. This aligns with the cultural norm of kuuki wo yomu (reading the air), in which unspoken rules serve as gatekeepers in settings such as banks or clinics (Nippoda, 2012). Consequently, for short-term students lacking long-term social capital, the inability to decode these "implicit norms" creates a functional barrier that necessitates technological support beyond literal translation.

DIGITAL COPING STRATEGIES AND THEIR LIMITATIONS

Although many Japanese universities have established formal support systems, these resources are often difficult to use in immediate, everyday situations and therefore underutilized (Masuda et al., 2005; Mojaverian et al.,

2013; Yeh et al., 2001). As a result, international students tend to rely on readily accessible alternatives, including peer networks, personal experience, and digital tools, to navigate daily challenges (J. S. Lee, 2017). Prior research further suggests that digital engagement often serves as the first point of support in the adaptation process, helping students take immediate action and make decisions in everyday situations. Such patterns can be understood as practical coping strategies, where digital tools often serve as the first point of support before direct real-world interaction (Xin et al., 2025).

However, these strategies remain limited in supporting the development of functional and interactional competence required for effective sociocultural adaptation. While social media platforms such as WeChat or Facebook provide essential emotional connection and identity negotiation (Forbush & Foucault-Welles, 2016; Pang, 2020; Zhou & Yin, 2024), they can exacerbate "information asymmetry" in the Japanese context. Wu and Ishii (2025) noted that the lack of explicit institutional criteria forces students into "herd mentality," relying on anecdotal peer advice that may be outdated or contextually irrelevant. Consequently, while this reliance validates and soothes emotions, it often fails to provide substantive guidance regarding specific and highly localized procedural rules.

Similarly, widely used translation applications serve as primary survival tools but exhibit critical functional limitations. While accessible, these tools often fail in high-pressure, context-dependent situations (Bradley & Al-Sabbagh, 2022; Liebling et al., 2020). Research indicates that they prioritize literal semantic accuracy over pragmatic competence (Liebling et al., 2020), frequently failing to convey the complex politeness levels (keigo) required in Japanese service encounters. Consequently, while these digital strategies provide partial support, they do not adequately foster the interactional competence required to navigate the implicit scripts of Japanese institutions.

Navigating Cultural Uncertainty in the AI Age: The Mediating Role of LLMs

In culturally ambiguous and high-uncertainty environments, individuals rely on sensemaking to interpret unfamiliar situations. As Weick (1995) defined, sensemaking is the process by which people give meaning to experience, particularly in ambiguous contexts. This process is especially critical in navigating the "information asymmetry" and "information silos" prevalent in Japan, as highlighted by Wu and Ishii (2025). In such settings, individuals must go beyond literal understanding to infer underlying institutional logic and appropriate actions.

In this context, LLMs may serve as context-aware mediating tools that support sensemaking by explaining underlying institutional procedures, such as the necessity of specific steps or requirements. By doing so, these systems have the potential to extend beyond translation and assist users in deriving meaning from culturally embedded situations. Recent scholarship has begun to operationalize this potential. Beyond academic learning, LLMs are increasingly

applied to lower psychological barriers to help-seeking (Mekni, 2021; Merikko & Silvola, 2024) and provide context-specific guidance via retrieval-augmented generation (RAG) technologies (Saha & Saha, 2024). Furthermore, emerging research suggests that AI agents can support identity negotiation, allowing young migrants to define their own cultural identities rather than conforming to pre-set categories (S. Lee et al., 2025). These studies collectively suggest that LLMs may support both functional and psychological aspects of adaptation.

However, the effectiveness of such support depends critically on technology trust. As Darmu'in et al. (2025) argued using Technology Trust Theory, users' willingness to rely on AI for high-stakes normative guidance is contingent on perceived reliability and validity. This trust is fragile; scholars caution that current LLMs often reflect Western-centric biases or lack specific cultural knowledge models, potentially limiting their effectiveness in high-context non-Western societies such as Japan (Adilazuarda et al., 2024; Ge et al., 2024). Consequently, there remains a lack of empirical evidence regarding whether LLMs can effectively bridge these "cultural blind spots" in real-world scenarios. This study addresses this gap by evaluating the performance of LLM tools in specific, culturally embedded tasks.

METHODOLOGY

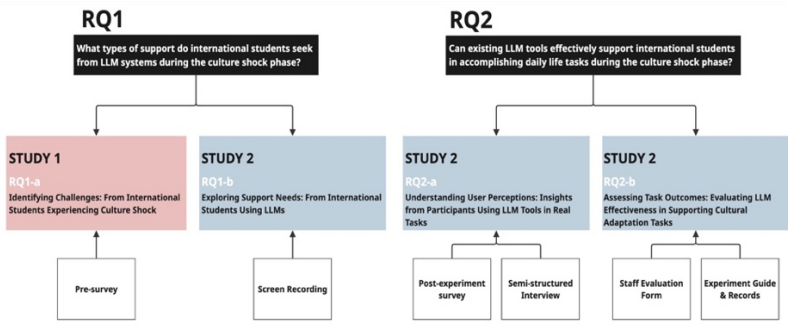

Figure 1: Research design linking two studies to RQ1 and RQ2. Study 1 explores initial challenges; Study 2 evaluates LLM support needs, user perceptions, and task effectiveness.

Research Design

This study adopts a sequential mixed-methods design conducted in two phases (Study 1 and Study 2), each corresponding to sub-questions derived from the main research questions (see Figure 1). Study 1 (Preliminary Survey; $N = 103$) addresses RQ1-a by identifying key challenges faced by international students during the culture shock phase. These findings informed the design of the task scenarios used in Study 2. Study 2 (User Study; $N = 40$) constitutes the primary empirical phase and is structured into two components corresponding to RQ1-b and RQ2. For RQ1-b, a sequential integration approach was applied, in

which screen recordings were treated as the primary data source to examine help-seeking behaviors, with task scenarios providing contextual grounding. For RQ2-a, a concurrent triangulation design (Bekhet & Zauszniewski, 2012) was employed, where post-task surveys captured participants' perceived experiences and semi-structured interviews provided explanatory insights into user behavior. For RQ2-b, quantitative measures, including staff evaluations and task completion records, were used to assess observable task outcomes across tools. Together, these components were integrated to examine how behavioral needs (RQ1) and the relationship between perception and performance (RQ2) characterize LLM-supported task engagement during the culture shock phase. Both studies received approval from the Research Ethics Committee of the Faculty of Library, Information and Media Science at the University of Tsukuba.

Study 1

Participants and Experiment Design

Participant demographics are summarized in Table 1. To identify key challenges faced during the culture shock phase, we conducted a pre-survey among international students at the University of Tsukuba. A total of 103 valid responses were obtained from students of diverse nationalities, most of whom had arrived within the past six months (60 female, 43 male). Participants were recruited through a purposive and volunteer-based sampling approach using both online platforms (Facebook, WeChat, LINE) and offline channels within the university.

As this study served as an exploratory phase to identify common adaptation challenges and inform the design of Study 2 tasks, no a priori power analysis was conducted. Instead, the sample size was determined based on practical feasibility and recruitment accessibility within this hard-to-reach population. The final sample size ($N = 103$) is consistent with prior survey-based and exploratory studies on international student adaptation, which typically range from approximately $N = 70$ to $N = 200$ (Al Juboori et al., 2025; GulRaihan & Sandaran, 2018; Kuo & Roysircar, 2006; Przyłęcki, 2018), supporting the adequacy of the sample for descriptive and exploratory analysis in this study context.

The questionnaire consisted of multiple-choice and open-ended items covering six demographic variables: gender, age, nationality, academic status, length of stay in Japan, and prior overseas study experience. Additional measures included self-rated Japanese proficiency, language learning motivation, perceived adaptation challenges, and coping strategies. Participants were also asked about their willingness to join a follow-up experiment (Study 2), and those who agreed provided contact information for further communication.

Table 1: Demographic information of all participants ($N = 103$)

Category	Item	No	%	Item	No	%
Sex	Male	43	41.75	Female	60	58.25
Age	18–29	86	83.50	29+	17	16.50
Region	East Asia	75	72.82	Southeast Asia	8	7.77
	South Asia	6	5.83	Central Asia	3	2.91
	South America	4	3.88	North America	1	0.97
	Western Europe	2	1.94	Africa	3	2.91
	Oceania	1	0.97			
Status	Bachelor's student	8	7.77	Master's student	33	32.04
	PhD student	17	16.50	Research student	45	43.69
Study in JP	<1 month	25	24.27	1–3 months	12	11.65
	3–6 months	29	28.16	>6 months	37	35.92
Study Abroad	Yes	27	26.21	No	76	73.79

Data Analysis

(RQ1-a) Identifying Challenges: From International Students Experiencing Culture Shock

Quantitative data from the pre-survey were analyzed using descriptive statistics (frequency counts and percentages) to identify the most prevalent sociocultural adaptation challenges. Responses to the multiple-choice items regarding everyday difficulties were ranked by frequency to determine high-friction domains. These quantitative rankings served as the primary basis for selecting the experimental scenarios in Study 2, ensuring that the tasks reflected real-world barriers to adaptation rather than hypothetical situations.

Results and Experiment Scenario Selection for Study 2

(RQ1-a) Identifying Challenges: From Participants Experiencing Culture Shock

To identify key adaptation challenges during the culture shock phase, we analyzed responses from the preliminary survey (Study 1). Quantitative analysis of the survey responses revealed that adaptation challenges were concentrated in language communication and institutional procedures. Specifically, language communication was the most frequently reported difficulty ($n = 67$), followed by opening a bank account ($n = 61$) and making phone appointments ($n = 44$). Other major stressors included housing ($n = 39$), transportation ($n = 38$), and medical visits ($n = 38$).

Based on these rankings, four scenarios were prioritized for the Study 2 user experiment: (1) Opening a bank account, (2) Medical visit process, (3) Making a phone appointment (hair salon), and (4) Garbage classification. Notably, although housing ($n = 39$) and transportation ($n = 38$) ranked highly, they were excluded due to constraints on experimental feasibility and simulation fidelity. These scenarios involve spatial mobility and long-term processes that are difficult to reproduce in a controlled laboratory setting. In contrast, tasks such as medical visits and garbage classification were selected as they represent high-stakes interactions that can be realistically simulated within the study design.

Study 2

Participants

Forty participants were purposively recruited from the Phase 1 survey. Eligibility required arrival in Japan within the past six months, reported adaptation challenges, and current enrollment at the University of Tsukuba. Participants were assigned to four groups ($n = 10$ each) through stratified randomization based on gender, nationality, and length of stay to maintain balance across conditions. The final sample included 15 males and 25 females, with most aged 18–29 (87.5%). Although East Asian students represented the largest proportion, each condition retained national diversity due to the stratified assignment. Detailed demographics are shown in Table 2. Each group used a designated tool: Group A used GPT-4o, Group B used Claude, Group C used Gemini, and Group D used Google Translate (control). Participants were anonymized using alphanumeric codes linked to their group (e.g., A1–A10).

Table 2: Participant demographics by tool condition ($N = 40$)

Variable	Category	GPT-4o	Claude	Gemini	Google Translate
		n (%)	n (%)	n (%)	n (%)
Gender	Male	5 (50)	3 (30)	4 (40)	3 (30)
	Female	5 (50)	7 (70)	6 (60)	7 (70)
Age	18–29	8 (80)	10 (100)	8 (80)	10 (100)
	30+	2 (20)	0 (0)	2 (20)	0 (0)
Region	East Asia	8 (80)	7 (70)	8 (80)	9 (90)
	Other	2 (20)	3 (30)	2 (20)	1 (10)
Academic status	Bachelor's student	2 (20)	2 (20)	1 (10)	1 (10)
	Master's student	1 (10)	2 (20)	2 (20)	4 (40)
	PhD student	1 (10)	1 (10)	1 (10)	1 (10)
	Research student	6 (60)	5 (50)	6 (60)	4 (40)
Length of stay in Japan	<1 month	2 (20)	2 (20)	2 (20)	1 (10)
	1–3 months	1 (10)	4 (40)	2 (20)	2 (20)
	3–6 months	7 (70)	4 (40)	6 (60)	7 (70)
Prior study abroad	Yes	1 (10)	2 (20)	2 (20)	4 (40)
	No	9 (90)	8 (80)	8 (80)	6 (60)

Note. Values are shown as n (%) within each tool condition (each $n = 10$). Region was summarized as East Asia vs. Other to improve readability; "Other" includes Central Asia, North America, South America, and Western Europe.

Details of the Experimental Setting in Study 2

User Scenario Setting

Scenario 1 - Opening a Bank Account: Participants simulated opening a general deposit account in a Japanese bank. Using a fictional identity profile and a personal seal (inkan), they completed a standardized application form, reflecting the structured nature of institutional procedures.

Scenario 2 - Medical Visit Process: Participants completed a two-stage clinical consultation, including filling out an intake questionnaire and describing symptoms to a physician. This task required precise expression in a context with limited explicit guidance.

Scenario 3 - Making a Phone Appointment with a Hair Salon: Participants completed a phone reservation and an in-person consultation, selecting a time and communicating a preferred hairstyle. The task involved both procedural coordination and context-sensitive expression.

Scenario 4 - Garbage Classification: Participants categorized items into seven bins and as-signed collection days based on a municipal guide and schedule. This task reflects the complexity of localized and often implicit disposal rules.

Tools Used in User Scenarios

The study compared three state-of-the-art LLMs against a widely adopted machine translation baseline. Participants were assigned to one of four experimental conditions: Group A (GPT-4o), Group B (Claude 3.5 Sonnet), Group C (Gemini Advanced), and Group D (Google Translate).

To ensure ecological validity, participants interacted with the latest stable model versions available at the time of their session (late 2024 to early 2025). For the Claude condition, all participants interacted exclusively with the identical claude-3-5-sonnet-20241022 version, as no new iterations were released during the data collection window (Anthropic, 2026). For GPT-4o, while updates occurred, official specifications confirm that the differences between the iterations used (2024-08-06 and 2024-11-20) lie strictly in developer-side API customization rather than intrinsic inference capabilities (OpenAI, 2024). For the Gemini condition, participants used the commercial web interface, which involved a generational transition (from the 1.5 series to the 2.0 series) and dynamic routing between Pro and Flash architectures. While this introduces architectural variability, it accurately reflects the ecological reality of interacting with commercial AI services. Importantly, official benchmarks demonstrate performance continuity: the newer 2.0 Flash matches or exceeds 1.5 Pro in core logical reasoning and comprehension metrics (Google, 2024, 2026). Furthermore, all underlying iterations were production-ready releases rather than experimental versions. Thus, the baseline capability and interaction quality remained highly comparable across all participants.

All tools were operated under default system parameters without custom system prompts to replicate typical user experiences. Google Translate was

selected as the control condition due to its ubiquity as a primary translation tool among international students (Pham et al., 2022). During the tasks, participants were restricted to interacting solely with their assigned tool and were not allowed to open external web browsers or applications. For the conversational LLMs, default internal retrieval and web-search capabilities were enabled, allowing the AI to pull external information autonomously. In contrast, Google Translate functioned strictly as a direct translation tool.

User Scenario Study Procedure

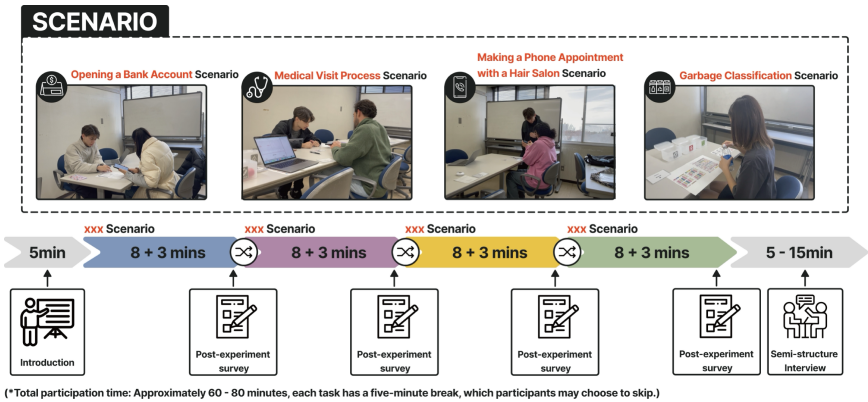

Figure 2: User Study Procedure. All participants will follow this procedure during the experiment. (The order of tasks will be determined by a random draw on the day of the experiment.)

An overview of the experimental procedure is shown in Figure 2. Sessions were conducted individually and lasted approximately 80 minutes. Upon arrival, participants received a standardized briefing via an instructional video introducing the assigned AI tool. Informed consent was obtained, including the right to withdraw at any time. To reduce order effects, the sequence of the four scenarios was randomized using a lottery method.

Each scenario simulated real-world pressure, consisting of a 3-minute preparation period followed by an 8-minute execution phase. Interlocutors (bank staff, doctors, or salon staff) followed standardized scripts to maintain consistency. A 5-minute break was provided between tasks to reduce fatigue.

To fully simulate natural, real-world interactions, participants were allowed to use their preferred languages when communicating with the tools. Additionally, follow-up prompting was completely unrestricted, meaning users could engage in multiple rounds of dialogue with the AI until they felt ready to proceed with the task.

Multi-modal data were collected through continuous screen and video recording of user–AI interactions. Task performance was evaluated immediately after each scenario by trained native Japanese raters using standardized rubrics.

The session concluded with a semi-structured interview examining perceived usefulness and cultural relevance. This component complemented the performance data by capturing subjective experience. Participants received a 1,060 JPY gift card as compensation.

Measures and Evaluation Instruments

Post-Experiment Survey

User perceptions were evaluated using a structured instrument developed specifically for this study, consisting of three distinct sub-scales.

Part 1 Task Experience: This 10-item sub-scale assessed the user's subjective experience during the interaction, covering dimensions of perceived effectiveness, confidence, and willingness to reuse. Items were rated on a 7-point Likert scale (1 = Strongly Disagree/Inefficient, 7 = Strongly Agree/Efficient).

Part 2 LLM Evaluation: Adapted from the CUC-FATE framework (Quttainah et al., 2024) and trust-oriented evaluation models (Liu et al., 2023), this 7-item sub-scale evaluated specific AI capabilities. Participants rated the tool on Reliability, Quality, Relevance, Fairness, Explainability, Human-likeness, and Translation Capability using a 5-point Likert scale (1= Very Slightly, 5 = Extremely).

Part 3 Interface Usability: This 9-item sub-scale measures the interaction design, focusing on ease of operation, function findability, and visual clarity. Items were rated on a 7-point Likert scale (1 = Very Difficult, 7 = Very Easy).

Staff Evaluation Form

To complement subjective evaluations, a single native Japanese staff member conducted and evaluated all interactive scenarios to eliminate interrater variability. Prior to the experiment, this staff member underwent an alignment session to internalize standardized rubrics and memorized standardized conversational scripts. During the sessions, the staff dynamically adapted to participants' reactions while strictly maintaining the fixed procedures and scripted core responses.

Each task was evaluated using two main criteria tailored to its nature: (1) language comprehension and overall task performance (with garbage classification including two sub-criteria: item sorting accuracy and correct placement of day-of-the-week cards) and (2) accuracy in written or physical outputs. Items were rated on a 0–5 Likert scale.

Because the staff member acted as both the interactive counterpart and the evaluator in these live, 1-on-1 scenarios, blinding to the tool conditions was not feasible. To minimize participant anxiety while ensuring evaluation accuracy, scoring was conducted discreetly immediately after each individual task, rather than at the end of the entire session, preventing any loss of observational detail due to memory decay. Furthermore, while statistical interrater consistency checks were not applicable due to the single-rater design, strict evaluation consistency

was enforced through explicit, objective decision rules. For example, a score of 5 denoted flawless completion with no staff intervention, while a score of 0 indicated critical failure to complete the task within the 8-minute limit.

Semi-Structured Interview

To triangulate the objective performance data and capture the nuances of user experience, a semi-structured interviews were conducted immediately after completion of the four tasks. Each session lasted approximately 5–15 minutes and was held individually. Interviews were carried out in Chinese or English according to participant preference to facilitate expression. All sessions were audio-recorded and transcribed verbatim for thematic coding. The interview protocol covered five thematic dimensions aimed at eliciting reflection on human–AI interaction:

Communication Experience: clarity of AI responses and specific moments of communication breakdown or success.

AI Performance in Contextual Tasks: evaluated the tool's ability to interpret cultural nuances and provide task-specific guidance.

Human-AI interaction: perceived partnership with AI and comparison to other support sources.

Satisfaction and Efficiency: perceived time-saving and overall task improvement.

Other Concerns: privacy, bias, cultural sensitivity, and over-reliance.

Data Analysis

(RQ1-b) Exploring Support Needs: From International Students Using LLMs

To quantify the specific types of support international students seek, we analyzed screen recordings from the 30 participants in the LLM groups. Distinct from traditional thematic analysis, which often focuses on the subjective interpretation of latent meanings, our approach prioritized the identification and quantification of explicit user support needs. Therefore, we employed a qualitative content analysis strategy (Elo & Kyngäs, 2008; Goddard & Melville, 2004).

The unit of analysis was defined as a single "interaction instance," representing any explicit user prompt or command directed at the AI tool. To ensure comprehensive coverage, events were operationalized into four structural types: (1) single-turn text questions, (2) image-based inquiries, (3) follow-up text prompts, and (4) image-based follow-ups. A total of 615 interaction instances were extracted. To ensure analytical rigor, an inductive coding process was employed. First, one researcher reviewed a randomly selected subset of the dataset (20%) to develop an initial coding scheme grounded in users' explicit interaction intents, such as requesting translation or seeking cultural context. This preliminary scheme was then refined through peer debriefing with a second

researcher, during which overlapping or related intents were systematically consolidated into nine distinct support categories. Following agreement on the finalized codebook, the primary researcher applied the scheme to the full dataset. Ambiguous cases were identified and resolved through consensus meetings to ensure coding consistency and minimize subjective bias.

(RQ2-a) Understanding User Perceptions: Insights from Participants Using LLM Tools in Real Tasks

To examine participants' perceptions and experiences when using different tools in culturally situated real-world tasks, we analyzed subjective data collected from post-experiment surveys and semi-structured interviews using a mixed-methods approach.

The post-experiment survey consisted of three sections assessing (1) Task Experience, (2) LLM Evaluation, and (3) Interface Usability. Descriptive statistics were calculated to summarize overall evaluation tendencies across tools. For multi-item dimensions, item scores were averaged to create composite measures. Internal consistency ranged from acceptable to good for exploratory constructs (Cronbach's $\alpha = .649 - .872$), including Ease of Use ($\alpha = .87$), Comfort and Learnability ($\alpha = .76$), and Operational Convenience ($\alpha = .65$). Missing responses (1.7%) were addressed using item-level median imputation.

For inferential analysis, given the small sample size per condition ($n = 10$) and the ordinal nature of Likert-scale data, non-parametric Kruskal–Wallis tests were used to compare tool conditions without assuming normality. When significant effects were detected, Dunn's post hoc tests with Bonferroni correction were applied to control for Type I error. As this study is exploratory, inferential results are interpreted as indicating overall patterns rather than definitive differences.

Interview transcripts were analyzed using inductive thematic analysis to capture participants' experiential accounts of tool use, including perceived usefulness, confidence, emotional reassurance, and moments of uncertainty. Coding proceeded iteratively, and emergent themes were reviewed across tool conditions.

Finally, quantitative and qualitative findings were integrated through triangulation, linking survey trends with participants' lived experiences to better understand how perceived tool qualities shaped interactions under cultural ambiguity.

(RQ2-b) Assessing Task Outcomes: Evaluating LLM Effectiveness in Supporting Cultural Adaptation Tasks

To evaluate objective task outcomes and identify sources of interactional difficulty in culturally situated tasks, we analyzed staff evaluation records, task execution logs, and structured task materials using a mixed analytic approach.

Staff evaluation forms provided quantitative ratings of participants' task performance across tools. Because evaluations were conducted for each task,

performance data were analyzed at the task level, with each scenario treated separately to capture scenario-specific performance differences. Descriptive statistics were first calculated to summarize overall trends, followed by non-parametric Kruskal–Wallis tests with Dunn's post hoc comparisons to explore potential differences across tool conditions. Effect sizes ($\epsilon 2$) were calculated to contextualize the magnitude of observed differences. As each participant contributed ratings to multiple tasks, the results are interpreted descriptively and with caution.

To capture practical task feasibility, task success rates were calculated based on execution records, defined as successful task completion within the allocated eight-minute time limit. These rates allowed comparison of tools' effectiveness in supporting timely task execution under realistic constraints.

To further examine the interactional demands embedded in different tasks, scripted staff prompts from three dialogue-based scenarios (banking procedures, medical intake, and phone appointment) were analyzed using structured content analysis (Elo & Kyngäs, 2008). Each prompt was coded along three dimensions: (1) openness (linguistic flexibility), (2) contextual dependence (degree of task-specific or institutional knowledge required), and (3) linguistic generation demands (extent of participant language production required). This classification enabled systematic comparison of interactional complexity across task types.

Finally, Pearson correlations were computed between participants' subjective task-experience ratings and staff-evaluated performance scores to explore the extent to which perceived usefulness and confidence aligned with observable task outcomes.

Result

To improve the clarity and coherence of the findings, the Results are organized to align with the study's central interpretive claims. We first present behavioral patterns derived from screen recordings to examine help-seeking needs (RQ1-b). We then report participants' subjective experiences (RQ2-a), followed by objective task outcomes based on staff evaluations and completion records (RQ2-b). This structure allows us to examine how behavioral needs, perceived support, and observed performance collectively inform the interpretation of LLM-supported adaptation. In doing so, the Results provide a more integrated understanding of how international students interact with LLMs and what types of support these systems offer in culturally unfamiliar contexts.

(RQ1-b) Exploring Support Needs: From International Students Using LLMs

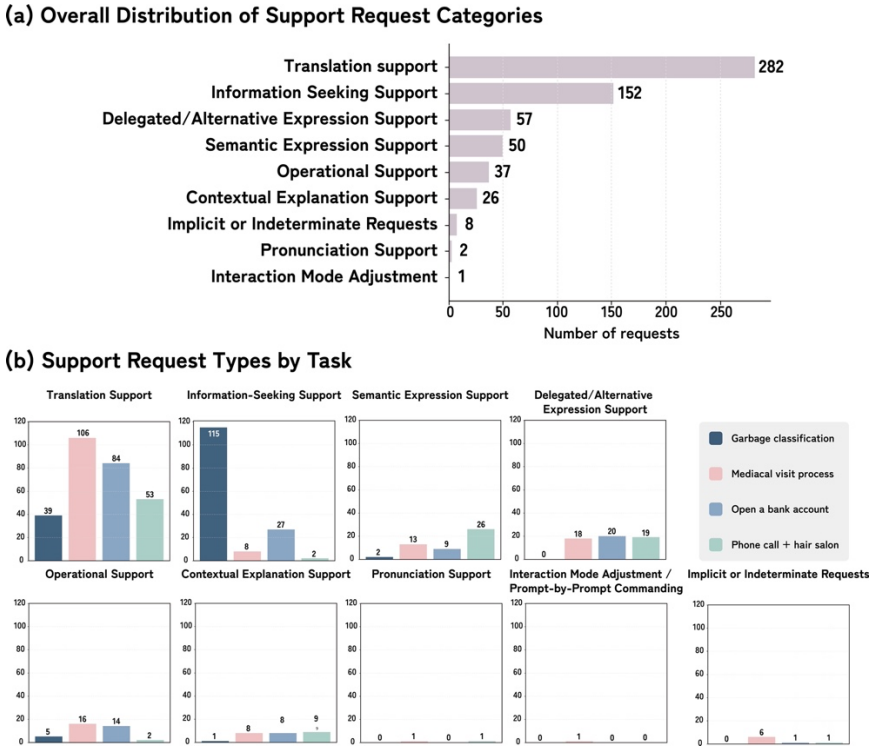

Figure 3: Distribution of LLM support requests across interaction categories. (a) Overall frequencies based on 615 interaction instances. (b) Distribution of support types across the four task scenarios.

To examine help-seeking behaviors in real-time interactions, we analyzed screen recordings as the primary data source. Analysis of 615 interaction instances extracted from screen recordings revealed that our participants' use of LLMs during cultural adaptation tasks extended well beyond simple translation. While Translation Support remained the most frequent category ($n = 282$), a substantial proportion of interactions focused on Information Seeking ($n = 152$) and Strategic Communication support, including Semantic Expression and Delegated Expression ($n = 107$) (Figure 3 (a)). This distribution indicates that participants relied on LLMs not only to convert language but also to interpret institutional rules and manage communicative challenges in unfamiliar cultural contexts, suggesting that, in this study, support needs extended beyond literal translation toward interpreting institutional expectations.

The nature of support sought was strongly task-dependent, shifting according to the interactional demands of each scenario (Figure 3 (b)). In procedural tasks such as garbage classification, support needs were predominantly informational ($n = 115$). Participants frequently queried local disposal rules (how to categorize specific items), treating the LLM as a cultural knowledge resource to make sense of implicit institutional logic. In contrast, tasks involving real-time interpersonal interaction, such as medical consultations and hair salon visits, elicited greater reliance on pragmatic expression support ($n = 39$, combined across tasks). In these contexts, participants sought assistance not only with vocabulary but also with sentence restructuring and appropriate politeness levels to conform to social norms, indicating a need to navigate interactional expectations in addition to linguistic content.

Notably, a pattern of Delegated/Alternative Expression Support was observed in high-interaction tasks ($n = 57$), where participants requested AI-generated utterances to be shown directly to staff. This behavior suggests that, under conditions of heightened uncertainty or anxiety, participants strategically used LLMs as communicative proxies to reduce the burden of direct verbal interaction, indicating that LLMs were used as interactional support to manage uncertainty rather than solely as translation tools. Together, these patterns suggest that during the culture shock phase, participants in this study used LLMs not merely as dictionaries but as adaptive resources for navigating both linguistic and institutional challenges. These patterns are further contextualized by participants' subjective perceptions and objective task outcomes discussed in RQ2.

(RQ2-a) Understanding User Perceptions: Insights from Participants Using LLM Tools in Real Tasks

Table 3: Descriptive statistics across all dimensions by tool condition

Dimension	GPT-4o	Claude	Gemini	Google Translate
Part 1: Task Experience				
Helpfulness	5.68 (0.91)	5.25 (1.11)	4.95 (1.53)	5.05 (1.25)
Satisfaction	5.25 (0.99)	5.35 (1.16)	5.10 (1.43)	5.35 (1.14)
Trust	5.93 (0.93)	6.20 (0.50)	5.53 (1.40)	5.60 (0.84)
Ease of Use	5.74 (1.10)	5.69 (0.99)	4.96 (1.52)	5.51 (0.58)
Confidence	5.43 (1.02)	5.43 (1.25)	5.10 (1.39)	5.63 (0.74)
Reuse Intention	6.05 (0.91)	5.25 (1.23)	4.75 (1.81)	5.85 (0.99)
Overall Value	5.98 (0.71)	5.53 (1.08)	5.13 (1.45)	5.70 (0.85)

Part 2: System Characteristics				
Reliability	4.08 (0.57)	4.10 (0.63)	3.70 (0.75)	3.83 (0.72)
Fairness	3.53 (1.35)	3.00 (1.38)	3.53 (1.23)	3.50 (0.87)
Quality	3.95 (0.54)	4.08 (0.54)	3.68 (0.67)	3.65 (0.57)
Relevance	4.28 (0.70)	4.33 (0.58)	3.75 (0.90)	3.88 (0.59)
Explainability	4.40 (0.57)	4.30 (0.50)	3.63 (0.93)	3.48 (0.83)
Translate Capability	4.43 (0.57)	4.30 (0.70)	4.28 (0.80)	3.88 (0.57)
Human-like	3.53 (1.06)	3.50 (1.06)	3.45 (1.06)	—
Part 3: Interface Usability				
Overall Ease of Use	6.13 (1.04)	6.00 (1.03)	5.25 (1.35)	6.15 (0.70)
Function Findability	6.00 (0.80)	5.98 (0.97)	5.30 (1.25)	5.90 (0.86)
Operational Convenience	5.74 (0.97)	6.00 (0.50)	5.66 (1.10)	5.54 (0.59)
Visual Layout	5.78 (1.16)	5.90 (0.84)	5.30 (1.35)	5.63 (0.67)
Comfort and Learnability	5.93 (0.87)	5.90 (0.86)	5.55 (1.19)	6.15 (0.64)

Note. Values are presented as M (SD); $N = 40$ ($n = 10$ per condition). Part 1 and Part 3 used 7-point Likert scales, and Part 2 used a 5-point scale. Asterisks (*) indicate dimensions with significant differences (none observed in this study). Detailed inferential statistics are reported in the main text.

To understand participants' subjective experiences when using LLM tools, we analyzed post-task survey responses and interview data. Building on the interaction patterns identified in RQ1, participants generally reported positive experiences in culturally situated tasks, with variation across tools in perceived trust and willingness to reuse (see Table 3).

Non-parametric Kruskal–Wallis tests were conducted to compare tool conditions across the survey dimensions. After aggregating responses to the participant level, no statistically significant between-tool differences were observed (all $p > .05$). For example, reuse intention showed descriptive variation across tools, but this difference was not statistically significant ($H = 3.96, p = .266$). These results suggest that subjective evaluations varied across tools at a descriptive level but did not differ reliably under the current experimental conditions. These findings align with participants' qualitative accounts, indicating strong acceptance of conversational LLMs for real-world use.

One participant noted that GPT-4o made communication with Japanese staff smoother and less stressful, explaining that "when I didn't understand something, I could immediately ask the system" (A3). Another emphasized the tool's value

as a substitute for unavailable social support during everyday encounters, stating that "since I can't always have a person with me, it's very valuable to have this kind of tool to help me do the task" (A10). At a descriptive level, Claude was associated with relatively high perceived trust ($M = 6.20$) and helpfulness ($M = 5.25$), though participants occasionally reported difficulty filtering excessive information. In contrast, Gemini tended to show lower reuse intention ($M = 4.75$) and was associated with difficulties in maintaining contextual continuity, reflecting more hesitant adoption.

Conversational LLMs were consistently evaluated not only on translation quality but also on interactional qualities such as explainability and contextual coherence. At a descriptive level, GPT-4o showed relatively higher ratings across perceived quality, explainability, and human-likeness (Human-like: $M = 3.53$), with participants frequently describing its responses as reliable and logically consistent. As one participant explained, "I don't care what language I input, I always trust the answer" (A9). Claude occupied an intermediate position: while its explanatory style was often perceived as clear and accessible, some participants noted information overload, remarking that "it lists a lot of information, but only a small part of it is actually useful" (B1). These qualitative accounts align with Gemini's consistently lower perceived capability scores across dimensions.

Despite these differences in perceived intelligence and interaction quality, overall usability ratings were high across all tools, including traditional translation systems. Google Translate received a high ease-of-use rating ($M = 6.15$), reflecting its familiarity and low learning cost. However, participants frequently distinguished surface-level usability from deeper task support. While the tool was described as simple to operate, it was often perceived as insufficient for complex, context-dependent interactions. One participant noted that "the problem stops at name translation, this hurdle cannot be crossed" (D1), highlighting a perceived ceiling in functionality. Taken together, the RQ2-a results suggest that conversational LLMs, particularly GPT-4o, were perceived as more supportive at a descriptive level, suggesting that perceived usefulness was closely tied to interactional reassurance rather than objective task performance, whereas traditional translation tools were valued primarily for efficiency and familiarity.

(RQ2-b) Assessing Task Outcomes: Evaluating LLM Effectiveness in Supporting Cultural Adaptation Tasks

To complement these subjective perceptions identified in RQ2-a, we further examined objective task outcomes.

Staff Evaluation Scores: Comparable Task Execution Across Tools

Kruskal–Wallis tests were conducted to examine whether staff-rated task performance differed across tools. Across the haircut, banking, and medical tasks, no significant differences were found for overall flow of communication or document completion (all $p > 0.50$). Effect sizes were negligible ($\epsilon 2 \approx 0$),

indicating minimal variation in observable performance across conditions, suggesting that objective task performance did not differ substantially across tools, despite descriptive variation observed in user perceptions.

A marginal trend appeared only in the garbage classification task for card placement flow ($H = 7.45, p = 0.059, \epsilon 2 = 0.12$). Although not statistically significant, this may indicate potential variability in how participants handled structured procedural steps. Overall, staff ratings indicate comparable levels of task execution across tools.

Task Completion Rates: Variation Driven by Procedural Complexity

Completion rates varied more by task than by tool. The haircut task showed the highest success, with GPT-4o, Claude, and Google Translate at 100% and Gemini at 80%. Garbage classification was also relatively high (GPT-4o: 100%; Claude and Gemini: 80%; Google Translate: 90%).

The medical visit process was the most difficult, with completion rates ranging from 30% (GPT-4o, Claude) to 60% (Google Translate), while Gemini reached 50%. Banking outcomes were moderate, with most tools between 60% and 70%. These results suggest that procedural difficulty within the time constraint influenced completion outcomes, indicating that task success depended more on navigating procedural and interactional demands than on differences between tools.

Interactional Complexity: A Source of Task Difficulty

To contextualize completion differences, we analyzed prompt structures across tasks. The medical scenario contained the highest level of open-ended and context-dependent interaction, requiring real-time symptom description. In contrast, banking relied on more structured factual exchanges, haircut appointments benefited from familiar vocabulary, and garbage classification reduced ambiguity through visible cues. These variations in interactional complexity likely contributed to differences in completion rates, highlighting how interactional complexity shaped task difficulty and constrained the effectiveness of LLM support.

Perceived vs. Objective Performance: Partial Alignment Across Tasks

Pearson correlation analysis between participants' average survey ratings and staff evaluation scores showed moderate positive correlations in most tasks. The strongest was in the medical consultation task for "Form-filling support" ($r = 0.54$, $p < 0.05$), and the weakest was in garbage sorting ($r = 0.07$), suggesting that alignment between subjective and objective performance may depend on task type, particularly interaction intensity and language demands.

Together, these findings suggest that differences in perceived usefulness are not necessarily aligned with observable task performance, highlighting a gap between subjective experience and objective outcomes.

FINDINGS AND DISCUSSION

This study examines how international students engage with LLMs during the culture shock phase and what types of support these systems provide in culturally unfamiliar environments. Based on observations from this study, the findings can be summarized into three central insights. First, participants in this study required institutional sensemaking rather than literal translation when navigating unfamiliar systems. Second, trust in LLMs appears to be shaped more by interactional reassurance than by objective task performance. Third, LLMs functioned as temporary scaffolds during early-stage adaptation rather than fully replacing institutional competence. Together, these findings suggest that while LLMs help reduce interactional uncertainty, they remain limited in supporting users through institutionally structured procedures that require situated understanding.

Institutional Sensemaking Beyond Literal Translation

Even participants with sufficient Japanese proficiency struggled when tasks required understanding institutional logic rather than vocabulary. Students could read the words but remained uncertain about what action was expected. Screen-record analyses (615 instances) repeatedly showed hesitation around items such as the Reiwa calendar (18 occurrences) and banking procedures such as auto-transfer thresholds. These were not translation failures but interpretation failures. Participants were effectively asking: What am I supposed to do here?

These patterns suggest a breakdown in sensemaking, where users are able to decode linguistic content but struggle to interpret its procedural implications. This challenge is particularly pronounced in high-context environments such as Japan, where institutional expectations are often implicit and require inference beyond what is explicitly stated. When LLMs provided literal explanations without clarifying implications or next steps, uncertainty persisted, sometimes leading to stalled interaction or incorrect entries. This gap illustrates a core cultural blind spot where linguistic comprehension does not translate into procedural confidence. In terms of sociocultural adaptation, this reflects a gap in functional and interactional competence: users may understand surface-level meaning but remain unable to translate that understanding into appropriate action within institutional contexts. Prior research similarly warns that current LLMs often remain at the level of surface cultural rendering rather than institutional reasoning (Adilazuarda et al., 2024; Singh et al., 2024). Our findings extend this argument by demonstrating how such limitations manifest in real-time adaptation tasks.

Trust as Interactional Reassurance Rather Than Performance Superiority

"I don't care what language I input. I always trust the answer." (A9) Despite limited statistical differences in objective performance across tools (many $p > 0.50$), at a descriptive level, GPT-4o tended to show higher willingness for reuse and perceived reliability, while Google Translate achieved comparable procedural

success but was sometimes evaluated less favorably. From a sociocultural adaptation perspective, trust in LLMs appears to be driven less by task performance than by the system's ability to provide interactional reassurance under conditions of uncertainty.

Participants' accounts indicate that this reassurance is tied to the social risks of communication in unfamiliar institutional settings, particularly in high-context environments (kuuki wo yomu). Some participants reported hesitation in asking others for help due to concerns about asking "obvious" questions or imposing on others (C6), whereas LLMs allowed them to ask freely without pressure.

In practice, participants often used LLMs to test phrasing, confirm understanding, or generate responses before acting. This suggests that LLMs function as a form of rehearsal that reduces the perceived cost of making mistakes. As a result, trust emerges from lowering the threshold for action. By reducing perceived social risk, LLMs enable users to engage in situations they might otherwise avoid, even when objective performance gains are limited.

Prior work similarly shows that AI trust is shaped by communicative style and perceived responsiveness rather than correctness alone (Afroogh et al., 2024; Gkinko & Elbanna, 2023; Ma et al., 2023; Wang et al., 2024). This also suggests that LLMs may partially substitute for immediate institutional support in situations where formal channels are perceived as in-accessible or effortful. In terms of sociocultural adaptation, this highlights the importance of interactional reassurance in facilitating early-stage engagement with unfamiliar institutional environments.

LLMs as Transitional Scaffolds in Sociocultural Adaptation

Beyond translation, participants frequently treated LLMs as interactional buffers that reduced the psychological burden of speaking in unfamiliar institutional settings. A recurrent strategy observed in the screen recordings (57 instances) involved showing AI-generated responses directly to staff, effectively delegating verbal responsibility to the system. Rather than asking only what a word means, users sought reassurance about how to act.

This behavior became particularly visible in high-interaction contexts such as medical visits and service encounters, where tone and politeness were perceived as high risk. One participant explained, "If I were speaking with a real Japanese person, I would be more nervous. With the AI, I can just ask" (A3). Another worried that incorrect phrasing might appear rude (B3). In such moments, LLMs functioned less as information providers and more as temporary scaffolds that stabilized participation under conditions of uncertainty.

At the same time, expectations shifted in low-interaction tasks. During garbage classification, participants often preferred speed and minimal output, sometimes expressing frustration when responses became overly elaborate. These contrasts indicate that students appeared to recalibrate what counted as "good support" depending on interpersonal exposure and perceived stakes.

This pattern suggests a shift toward mediated and lower-risk forms of engagement, consistent with prior work showing that individuals in cross-cultural

contexts often avoid direct or formal help-seeking and instead rely on informal or technologically mediated alternatives (Masuda et al., 2005; Mojaverian et al., 2013; Pang, 2020). Importantly, these findings suggest that LLMs do not replace users' need to develop sociocultural competence but instead support participation before such competence is fully established, functioning as transitional scaffolds between initial uncertainty and independent performance.

While these findings provide insight into LLM-supported adaptation during the culture shock phase, they should be interpreted within the scope of this study. The sample consisted primarily of recently arrived international students at a single institution, with a notable concentration of Chinese-speaking participants. This composition may have shaped observed interaction patterns, particularly in relation to translation strategies, perceptions of institutional difficulty, and trust in LLM-mediated support.

Accordingly, the results do not aim to generalize across all international student populations but rather to illuminate how LLMs are used and experienced in early-stage, high-uncertainty contexts. Future work should examine more diverse populations and settings to further validate and extend these patterns.

FROM FINDINGS TO DESIGN IMPLICATIONS

The findings indicate that linguistic accuracy alone is insufficient to support international students during the culture shock phase. Breakdowns emerged primarily at the level of procedural meaning rather than vocabulary, as participants often understood translations but remained uncertain about appropriate next actions. This suggests that LLM-based support should move beyond semantic rendering toward clarifying institutional expectations and locally normative pathways. The divergence between subjective trust and objective success highlights the importance of interactional experience. Participants consistently preferred systems that provided coherence, predictability, and reassurance, suggesting that effective support involves not only correctness but also the reduction of interactional anxiety.

Participants also positioned LLMs as transitional partners that support rehearsal, confirmation, and temporary delegation, thereby scaffolding gradual participation in unfamiliar environments. At the same time, support expectations varied by task structure: interpersonal scenarios required pragmatic sensitivity, whereas procedural contexts prioritized efficiency.

Together, these findings suggest that future systems should move beyond translation-centered assistance toward task-adaptive models that support sensemaking, confidence, and situated participation.

LIMITATIONS

This study explored how LLMs may support international students during the culture shock phase in this context, revealing needs that extend beyond language assistance. Nevertheless, several limitations should be acknowledged.

First, the study was conducted at a single institution with a relatively small sample size. While the design prioritized in-depth observation of situated behaviors during complex real-world tasks, the number of participants limits the generalizability of the findings. In addition, most participants were Chinese-speaking students. This distribution reflects both national and local enrollment patterns (Japan Student Services Organization (JASSO), 2025; University of Tsukuba, 2024b), yet the findings remain grounded in a relatively specific linguistic and cultural context. Interaction norms, tone expectations, and institutional familiarity may vary across student populations, which may in turn shape how LLMs are used and interpreted in practice. Future research should therefore validate these findings with larger and more diverse cohorts.

Second, the experiment focused on short-term, scenario-based activities to capture immediate adaptation challenges. Although this approach enabled close examination of real-time decisions and emotional reactions, it does not reveal how trust, strategies, or reliance on AI may evolve over time. Longitudinal investigations are needed to understand how relationships with LLM tools develop beyond initial encounters.

Finally, we relied on commercially available systems and had no access to internal algorithms or update mechanisms. Many processes shaping interaction, such as tone modulation or con-textual inference, remain opaque. This black-box condition restricts our ability to identify which system components influence outcomes, highlighting a broader methodological challenge in evaluating real-world AI services. In addition, variations in model versions across the study period may have introduced further variability in interaction outcomes.

CONCLUSION AND FUTURE WORK

The culture shock phase is often when international students need support most, yet appropriate assistance remains difficult to access. Through four real-world tasks, this study examined how existing LLM tools operate within this critical period, providing an empirical account of how AI becomes embedded in situated adaptation practices.

In this study, difficulties arose less from vocabulary than from limited understanding of institutional logic, procedural expectations, and culturally appropriate interaction styles. Even when translations were accurate, participants often remained uncertain about what actions should follow. Meanwhile, systems that offered natural dialogue and coherent explanations were highly valued, despite showing no clear advantage in objective completion rates. This divergence suggests that, during early-stage adaptation, participants in this study appeared to seek reassurance and interpretive support as much as practical solutions.

These findings suggest that LLMs may function not only as language tools but also as sense-making partners that help users manage ambiguity, regulate tone, and sustain confidence. At the same time, recurring breakdowns around institutional knowledge highlight the limitations of current systems in converting surface understanding into procedural certainty.

Future research should move beyond short-term task success toward longer-term adaptation trajectories. Larger and more diverse samples, as well as longitudinal approaches, are needed to understand how trust evolves and whether repeated interaction integrates AI into students' broader support ecologies.

As AI technologies increasingly serve as everyday companions, evaluation criteria may need to expand beyond correctness to include how systems structure understanding, stabilize emotions, and guide action under uncertainty. By situating LLM use within the context of international students' early-stage adaptation, this study highlights how AI-mediated support reshapes how students navigate institutional uncertainty, offering new insight into the role of technology in sociocultural adaptation.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT only for language refinement (e.g., grammar and clarity). All conceptual contributions, study design, data analysis, and interpretation were generated and developed by the authors. The final manuscript was reviewed, edited, and approved by the authors, who take full responsibility for its content.

Acknowledgment

In the preparation of this manuscript, we utilized artificial intelligence (AI) tools for content creation in the following capacity:

☐ *None*
✓ *Some sections, with minimal or no editing*
☐ *Some sections, with extensive editing*
☐ *Entire work, with minimal or no editing*
☐ *Entire work, with extensive editing*

REFERENCES

Adilazuarda, M. F., Mukherjee, S., Lavania, P., Singh, S. S., Aji, A. F., O'Neill, J., Modi, A., & Choudhury, M. (2024, November). Towards measuring and modeling "culture" in LLMs: A survey. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Proceedings of the 2024 conference on empirical methods in natural language processing* (pp. 15763–15784). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.882

Afroogh, S., Akbari, A., Malone, E., Kargar, M., & Alambeigi, H. (2024). Trust in AI: Progress, challenges, and future directions. *Humanities and Social Sciences Communications*, 11, 1568. https://doi.org/10.1057/s41599-024-04044-8

Al Juboori, R., Barker, D., & Kim, Y. J. (2025). Predictors of academic adjustment among international students in rural southern USA.

International Journal of Environmental Research and Public Health, 22(2), 253. https://doi.org/10.3390/ijerph22020253

Anthropic. (2026). Model deprecations [Accessed: 2026-04-14]. https://platform.claude.com/docs/en/about-claude/model-deprecations

Asian Students Cultural Association (ABK) & Benesse Corporation. (n.d.). Japan study support (JPSS) [Accessed: 2025-06-08]. https://www.jpss.jp/en/

Bekhet, A. K., & Zauszniewski, J. A. (2012). Methodological triangulation: An approach to understanding data. *Nurse Researcher*, 20(2), 40–43. https://doi.org/10.7748/nr2012.11.20.2.40.c9442

Bradley, L., & Al-Sabbagh, K. W. (2022). Mobile language learning designs and contexts for newly arrived migrants. *Australian Journal of Applied Linguistics, 5*(3), 179–189. https://doi.org/10.29140/ajal.v5n3.53si5

Colace, F., De Santo, M., Lombardi, M., Pascale, F., Pietrosanto, A., & Lemma, S. (2018). Chatbot for e-learning: A case of study. *International Journal of Mechanical Engineering and Robotics Research*, 7(5), 528–533. https://doi.org/10.18178/ijmerr.7.5.528-533

Darmu'in, D., Nasikhin, Darnoto, Sofanudin, A., Maarif, M. A., & Haryanto, J. T. (2025). Perceived barriers to ChatGPT integration in Islamic studies education: Insights from Malaysian international students in Indonesia. *Journal of International Students*, 15(11), 141-164. https://doi.org/10.32674/h99a4p06

Elo, S., & Kyngäs, H. (2008). The qualitative content analysis process. *Journal of Advanced Nursing*, 62(1), 107–115. https://doi.org/10.1111/j.1365-2648.2007.04569.x

Forbush, E., & Foucault-Welles, B. (2016). Social media use and adaptation among chinese students beginning to study in the United States. *International Journal of Intercultural Relations, 50*, 1–12. https://doi.org/10.1016/j.ijintrel.2015.10.007

Ge, X., Xu, C., Misaki, D., Markus, H. R., & Tsai, J. L. (2024). How culture shapes what people want from AI. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems(CHI '24, Article 95, pp. 1–15)*. Association for Computing Machinery. https://doi.org/10.1145/3613904.3642660

Gkinko, L., & Elbanna, A. (2023). Designing trust: The formation of employees' trust in conversational ai in the digital workplace. *Journal of Business Research, 158*, 113707. https://doi.org/10.1016/j.jbusres.2023.113707

Goddard, W., & Melville, S. (2004). *Research methodology: An introduction* (2nd). *Juta and Company Ltd*.

Google. (2024). Updated gemini models: Reduced 1.5 pro pricing, increased rate limits, and more [Accessed: 2026-04-14]. https://developers.googleblog.com/en/updated-gemini-models-reduced-15-pro-pricing-increased-rate-limits-and-more/

Google. (2026). Gemini api: Models [Accessed: 2026-04-14]. https://ai.google.dev/gemini-api/docs/models

GulRaihan, M., & Sandaran, S. (2018). Sociocultural adaptation challenges of international students at a higher learning institution in malaysia. *LSP International Journal,* 4(2), 85-101. https://doi.org/10.11113/lspi.v4n2.58

Japan Student Services Organization. (2024). Result of International Student Survey in Japan, 2023 [Accessed: 2025-06-08]. https://www.studyinjapan.go.jp/en/statistics/enrollment/data/2405241100.html

Japan Student Services Organization (JASSO). (2025). Result of International Student Survey in Japan, 2024 [Accessed: 2025-08-12]. https://www.studyinjapan.go.jp/en/_mt/2025/04/data2024z_e.pdf

Jayasinghe, R. P. C. K., & Rathnayake, R. B. P. M. (2022). The Adaptability of Postgraduate Students to a Foreign Environment with Particular Reference to the Japanese Cultural Context. *Sri Lankan Journal of Management*, 27(2), 133–157. https://doi.org/10.33939/SLJM.27.02.05.2022

Kamalov, F., Santandreu Calonge, D., & Gurrib, I. (2023). New era of artificial intelligence in education: Towards a sustainable multifaceted revolution. *Sustainability*, 15(16), 12451. https://doi.org/10.3390/su151612451

Kuo, B. C., & Roysircar, G. (2006). An exploratory study of cross-cultural adaptation of adolescent taiwanese unaccompanied sojourners in canada. *International Journal of Intercultural Relations*, 30(2), 159–183. https://doi.org/10.1016/j.ijintrel.2005.07.007

Lee, J. S. (2017). Challenges of international students in a japanese university: Ethnographic perspectives. *Journal of International Students*, 7(1), 73–93. https://doi.org/10.32674/jis.v7i1.246

Lee, S., Choi, D., Truong, L., Sawhney, N., & Paakki, H. (2025). Into the unknown: Leverag-ing conversational ai in supporting young migrants' journeys towards cultural adaptation. *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3706598.3713091

Liebling, D. J., Lahav, M., Evans, A., Donsbach, A., Holbrook, J., Smus, B., & Boran, L. (2020). Unmet needs and opportunities for mobile translation ai. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–13. https://doi.org/10.1145/3313831.3376261

Liu, Y., Yao, Y., Ton, J.-F., Zhang, X., Guo, R., Cheng, H., Klochkov, Y., Taufiq, M. F., & Li, H. (2023). Trustworthy llms: A survey and guideline for evaluating large language models' alignment. https://arxiv.org/abs/2308.05374

Ma, S., Lei, Y., Wang, X., Zheng, C., Shi, C., Yin, M., & Ma, X. (2023). Who should i trust: Ai or myself? leveraging human and ai correctness likelihood to promote appropriate trust in AI-assisted decision-making. *Proceedings of the 2023 CHI Conference on Hu-man Factors in Computing Systems*. https://doi.org/10.1145/3544548.3581058

Masgoret, A.-M. (2006). Examining the role of language attitudes and motivation on the sociocultural adjustment and the job performance of sojourners in spain. *International Journal of Intercultural Relations*, 30(3), 311–331. https://doi.org/10.1016/j.ijintrel.2005.08.004

Masuda, A., Suzumura, K., Beauchamp, K., Howells, G., & Clay, C. (2005). United states and japanese college students' attitudes toward seeking professional psychological help. *International Journal of Psychology*, 40(5), 303–313. https://doi.org/10.1080/00207590444000339

Mekni, M. (2021). An artificial intelligence based virtual assistant using conversational agents. *Journal of Software Engineering and Applications*, 14, 455–473. https://doi.org/10.4236/jsea.2021.149027

Merikko, J., & Silvola, A. (2024). An ai agent facilitating student help-seeking: Producing data on student support needs (M. Hlosta, I. Moser, & B. Flanagan, et al., Eds.), *Joint proceedings of lak 2024 workshops, co-located with 14th international conference on learning analytics and knowledge (lak 2024)* (pp. 185–194, Vol. 3667). CEUR-WS.org. https://ceur-ws.org/Vol-3667/

Mojaverian, T., Hashimoto, T., & Kim, H. (2013). Cultural differences in professional help seeking: A comparison of japan and the u.s. *Frontiers in Psychology*, 3, 615. https://doi.org/10.3389/fpsyg.2012.00615

Murphy-Shigematsu, S. (2002). Psychological barriers for international students in japan. *International Journal for the Advancement of Counselling*, 24(1), 19–30. https://doi.org/10.1023/A:1015076202649

Nikkei. (2023). Japan aims to be top study abroad destination in Asia again [Accessed: 2025-06-08]. https://www.nikkei.com/article/DGXZQOUA165GW0W3A310C2000000/

Nippoda, Y. (2012). Japanese students' experience of adaptation and acculturation of the united kingdom. *Online Readings in Psychology and Culture*, 8. https://doi.org/10.9707/2307-0919.1071

Oberg, K. (1960). Cultural shock: Adjustment to new cultural environments. *Practical Anthropology*, os-7(4), 177–182. https://doi.org/10.1177/009182966000700405

OpenAI. (2024). Gpt-4o model [Accessed: 2026-04-14]. https://developers.openai.com/api/docs/models/gpt-4o

Pang, H. (2020). Is active social media involvement associated with cross-culture adaption and academic integration among boundary-crossing students? *International Journal of Intercultural Relations*, 79, 71–81. https://doi.org/10.1016/j.ijintrel.2020.08.005

Park, H., & Ahn, D. (2024). The promise and peril of chatgpt in higher education: Oppor-tunities, challenges, and design implications. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3613904.3642785

Pham, A. T., Nguyen, Y. N. N., Tran, L. T., Huynh, K. D., Le, N. T. K., & Huynh, P. T. (2022). University students' perceptions on the use of google translate: Problems and solutions. *International Journal of Emerging Technologies in Learning*, 17(4), 79–94. https://doi.org/10.3991/ijet.v17i04.28179

Przyłęcki, P. (2018). International students at the medical university of Łódź: Adaptation challenges and culture shock experienced in a foreign country.

Central and Eastern European Migration Review, 7(2), 209–232. https://doi.org/10.17467/ceemr.2018.13

Quttainah, M., Mishra, V., Madakam, S., Lurie, Y., & Mark, S. (2024). Cost, usability, credibility, fairness, accountability, transparency, and explainability framework for safe and effective large language models in medical education: Narrative review and qualitative study. *JMIR AI*, 3, e51834. https://doi.org/10.2196/51834

Saha, B., & Saha, U. (2024). Enhancing international graduate student experience through ai-driven support systems: A llm and rag-based approach. *2024 International Con-ference on Data Science and Its Applications (ICoDSA)*, 300–304. https://doi.org/10.1109/ICoDSA62899.2024.10651944

Searle, W., & Ward, C. (1990). The prediction of psychological and sociocultural adjust-ment during cross-cultural transitions. *International Journal of Intercultural Relations*, 14(4), 449–464. https://doi.org/10.1016/0147-1767(90)90030-Z

Singh, P., Patidar, M., & Vig, L. (2024, November). Translating across cultures: LLMs for intralingual cultural adaptation. In L. Barak & M. Alikhani (Eds.), *Proceedings of the 28th conference on computational natural language learning* (pp. 400–418). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.conll-1.30

Study in Japan. (n.d.). Support program for international students [Accessed: 2025-06-08]. https://www.studyinjapan.go.jp/en/about/support-program.html

Sumer, S. (2009). International students' psychological and sociocultural adaptation in the united states [ERIC ED515598].

Sydoruk, T. (2024). How ai can reduce unemployment rate among vulnerable population of new immigrants: Assimilation issues resolved with ai. *Economics & Education*, 9(2), 20–25. https://doi.org/10.30525/2500-946X/2024-2-3

Tran, H. N., Inosaki, A., & Jin, C.-H. (2022). On campus support and satisfaction of international students: A review of japanese literature. *The IAFOR Conference on Ed-ucational Research & Innovation 2022 Official Conference Proceedings*, 1–16. https://doi.org/10.22492/issn.2435-1202.2022.1

University of Tsukuba. (2024a). Tutor system for international students [Accessed: 2025-07-22]. https://www.tsukuba.ac.jp/en/campuslife/support-international/tutor/

University of Tsukuba. (2024b). University of tsukuba factbook 2024 [Accessed: 2025-08-19]. https://www.tsukuba.ac.jp/about/factbook/factbook-full-ver-r6.pdf

Wang, J., Hu, H., Wang, Z., Yan, S., Sheng, Y., & He, D. (2024). Evaluating large language models on academic literature understanding and review: An empirical study among early-stage scholars. *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3613904.3641917

Ward, C., Bochner, S., & Furnham, A. (2001). *The psychology of culture shock* (2nd). Rout-ledge.

Ward, C., & Kennedy, A. (1999). The measurement of sociocultural adaptation. *International Journal of Intercultural Relations*, 23(4), 659–677. https://doi.org/10.1016/S0147-1767(99)00014-0

Ward, C., Okura, Y., Kennedy, A., & Kojima, T. (1998). The u-curve on trial: A longitudinal study of psychological and sociocultural adjustment during cross-cultural transition. *International Journal of Intercultural Relations*, 22(3), 277–291. https://doi.org/10.1016/S0147-1767(98)00008-X

Weick, K. E. (1995). Sensemaking in organizations. *Sage Publications Thousand Oaks*, CA.

Wu, Q., & Ishii, H. (2025). Decision strategies and influencing factors of international stu-dents' university entrance in japan. *Journal of International Students*, 15(10), 61–84. https://doi.org/10.32674/myc3w256

Xia, J. (2009). Analysis of impact of culture shock on individual psychology. *International Journal of Psychological Studies*, 1(2), 97–101. https://doi.org/10.5539/ijps.v1n2p97

Xin, Y., Shusheng, D., Weina, H., & Yan, D. (2025). How does digital connection shape cultural adaptation? the impact of social media use on cross-cultural adaptation of international students in China. *Journal of International Students*, 15(9), 1–26. https://doi.org/10.32674/0wkqv704

Yeh, C., Inose, M., Kobori, A., & Chang, T. (2001). Self and coping among college students in japan. *Journal of College Student Development*, 42, 242–256.

Yonezawa, A. (2020). Challenges of the japanese higher education amidst population decline and globalization. *Globalisation, Societies and Education*, 18(1), 43–52. https://doi.org/10.1080/14767724.2019.1690085

Zhou, S., & Yin, J. (2024). International students' social media use: An integrative review of research over a decade. *Journal of Studies in International Education*, 29(1), 42–63. https://doi.org/10.1177/10283153241275037

Author bios

ChunPi Hsieh is a master's student in the Graduate School of Comprehensive Human Sciences at the University of Tsukuba, Japan. Her research interests include AI-mediated cultural adaptation, Human–Computer Interaction (HCI), and support design for international students. Email: jocelyn85419@gmail.com

Yoichi Ochiai is a Professor at the Institute of Library, Information and Media Science and Tsukuba Institute for Advanced Research (TIAR), University of Tsukuba, Japan. His research spans kansei informatics, Human–Computer Interaction (HCI), virtual reality, human–AI collaboration, AI-driven accessibility, computational holography, ultrasonic user interfaces, laser-based user interfaces, and differential ontology. Email: wizard@slis.tsukuba.ac.jp

Ichiro Matsuda is a master's student in the Graduate School of Comprehensive Human Sciences at the University of Tsukuba, Japan. His work lies at the intersection of Human–Computer Interaction (HCI) and Human–Agent Interaction (HAI), focusing on systems in which humans design and observe autonomous debating agents. Email: ichi6m@digitalnature.slis.tsukuba.ac.jp

Tatsuki Fushimi is an Assistant Professor at the Institute of Library, Information and Media Science and Tsukuba Institute for Advanced Research (TIAR), University of Tsukuba, Japan. His research interests include applied physics and Human–Computer Interaction (HCI).
Email: tfushimi@digitalnature.slis.tsukuba.ac.jp

Jingjing Li is an Assistant Professor at the Institute of Library, Information and Media Science, University of Tsukuba, Japan. Her research focuses on Human–Computer Interaction (HCI) , user evaluation of digital media, digital exhibitions, and communication design. Email: li@digitalnature.slis.tsukuba.ac.jp