Quantifying the Impact of User Communication Diversity on LLM Agent Performance: A Framework with Information-Theoretic Decomposition

Anonymous Author(s)

ABSTRACT

Large language model (LLM) agents are increasingly deployed in task-oriented conversational settings, yet their robustness to the natural diversity of human communication remains poorly understood. Real users differ along dimensions including formality, verbosity, politeness norms, dialect, cultural context, and domain expertise—but how much does this variation affect whether an agent actually completes a task? We propose a framework for systematically quantifying this impact, built on three contributions: (1) a six-dimensional *Communication Style Space* grounded in sociolinguistic theory that parameterizes user diversity; (2) the *Communication Diversity Sensitivity Index* (CDSI), a scalar metric summarizing an agent's robustness to style variation; and (3) an information-theoretic decomposition that separates task outcome uncertainty into content-attributable and style-attributable components. In controlled experiments across 4 agent configurations, 12 user profiles, 4 task domains, and 19,200 simulated dialogues, we find that communication style accounts for 1.5%–7.5% of task success uncertainty, with dialect distance and cultural context as the most impactful axes ($\rho = -0.37$ and -0.36 for the most vulnerable agent). Agent CDSI scores range from 0.259 (robust) to 0.608 (highly sensitive), and all agents exhibit statistically significant performance disparities across demographic groups ($p < 10^{-18}$). Calibration gaps are largest for L2 speakers and high-context communicators, reaching 0.80 between confidence and actual success. These findings establish that communication diversity is a measurable and significant factor in agent performance and provide actionable metrics for auditing and improving equity.

ACM Reference Format:

Anonymous Author(s). 2026. Quantifying the Impact of User Communication Diversity on LLM Agent Performance: A Framework with Information-Theoretic Decomposition. In *Proceedings of ACM Conference (Conference'17)*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Task-oriented conversational agents powered by large language models (LLMs) are being deployed across customer service, technical support, healthcare, and education [12, 16]. These agents must parse user requests, extract relevant information, and complete tasks—all through natural language dialogue. Yet the users they serve are linguistically diverse: they vary in formality, verbosity, politeness conventions, dialect, cultural communication norms, and domain expertise.

A growing body of evidence suggests this diversity affects outcomes. Seshadri et al. [14] demonstrate that LLM-simulated users are unreliable proxies for real users in agentic evaluations, finding disparate success rates across dialects and age groups. They note that "users might vary along dimensions such as formality, verbosity, and politeness norms—but it remains unclear how much this diversity meaningfully impacts agent performance and task success" [15]. This observation identifies a critical open problem: we lack a quantitative framework for measuring and decomposing the impact of user communication diversity on agent task completion.

This gap matters for three reasons. First, *equity*: if agents systematically fail for users with non-standard communication styles, they perpetuate exclusion. Second, *evaluation validity*: benchmarks that collapse communication diversity into standardized instructions will overestimate real-world performance. Third, *design*: without understanding which dimensions of diversity drive failures, we cannot build targeted mitigations.

We address this open problem with three contributions:

(1) **Communication Style Space.** A six-dimensional parameterization of user communication diversity grounded in sociolinguistic theory (Brown and Levinson's politeness theory [3], Biber's register dimensions [1], Hall's cultural context framework [6], and the World Englishes paradigm [10]).
(2) **Communication Diversity Sensitivity Index (CDSI).** A metric quantifying how much an agent's task success rate degrades as user style deviates from the training-data norm, with per-axis decomposition and equity sub-metrics.
(3) **Information-theoretic decomposition.** A method for separating task outcome uncertainty into content-attributable (what was said) and style-attributable (how it was said) components, based on conditional mutual information.

In controlled experiments with 4 agent configurations, 12 sociolinguistic user profiles, 4 task domains, and 19,200 simulated dialogues, we find that communication style is a statistically significant predictor of task success for all agents tested (χ^2 tests: $p < 10^{-18}$), with CDSI scores ranging from 0.259 to 0.608 and style accounting for up to 7.5% of outcome uncertainty.

1.1 Related Work

Sociolinguistic variation in NLP.. Research on dialect robustness has demonstrated that NLP systems degrade on non-standard English [2, 17]. These studies focus primarily on classification and generation tasks rather than multi-turn agentic task completion. Danescu-Niculescu-Mizil et al. [5] provide computational operationalizations of politeness, which we build upon.

Conference'17, July 2017, Washington, DC, USA

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Task-oriented dialogue evaluation. Classical task-oriented dialogue benchmarks such as MultiWOZ [4] and DSTC [7] measure slot-filling accuracy and task success but use templated or crowd-sourced utterances that do not capture real communication diversity. Modern agent benchmarks like WebArena [16] and SWE-bench [8] evaluate complex capabilities but use standardized, clean instructions.

Robustness testing. Ribeiro et al. [13] introduce CheckList, a behavioral testing framework for NLP that includes linguistic perturbations. Our work extends this paradigm from classification to agentic task completion and from ad-hoc perturbations to theory-grounded sociolinguistic dimensions.

LLM user simulation. Seshadri et al. [14] show that LLM-simulated users diverge from real users in agentic evaluations and identify communication diversity as a key source of this gap. Joshi et al. [9] evaluate LLM persona fidelity. Our framework provides the quantitative methodology that these works identify as missing.

Positioning. No existing work systematically varies user communication style along multiple sociolinguistic dimensions and measures the causal impact on multi-turn agent task success with attribution to specific axes and failure modes. We fill this gap.

2 METHODS

2.1 Communication Style Space

We define a six-dimensional communication style space $\mathcal{S} = [0,1]^6$ where each axis represents a sociolinguistic dimension of user variation:

(1) **Formality** (s_1): Register from colloquial ($s_1 = 0$) to formal ($s_1 = 1$), following Biber's register dimensions [1].
(2) **Verbosity** (s_2): From terse single-clause utterances ($s_2 = 0$) to elaborate multi-sentence turns ($s_2 = 1$).
(3) **Politeness** (s_3): From direct/blunt ($s_3 = 0$) to heavily hedged and indirect ($s_3 = 1$), grounded in Brown and Levinson [3].
(4) **Dialect distance** (s_4): From Standard American English ($s_4 = 0$) to maximal dialect divergence ($s_4 = 1$), drawing on the World Englishes framework [10].
(5) **Cultural context** (s_5): From low-context/explicit ($s_5 = 0$) to high-context/implicit ($s_5 = 1$), following Hall [6].
(6) **Domain expertise** (s_6): From lay description ($s_6 = 0$) to expert jargon ($s_6 = 1$).

A user's communication style is a vector $\mathbf{s} = (s_1, \ldots, s_6) \in \mathcal{S}$. We define a *standard style* $\mathbf{s}_0 = (0.4, 0.4, 0.4, 0.0, 0.2, 0.3)$ representing the communicative norms most represented in LLM training data, and compute style distance as $d(\mathbf{s}) = \|\mathbf{s} - \mathbf{s}_0\|_2$.

2.2 User Profiles

We construct 12 canonical user profiles spanning the style space (Table 1), including a baseline profile at $\mathbf{s}_0$, style extremes (formal-verbose, casual-terse, high-politeness), dialect variants (AAVE, Indian English, L2 beginner), cultural variants (high-context), age-related styles (elderly, teen), and professional registers (expert, corporate). Each profile is assigned a demographic group label for equity analysis.

2.3 Task Scenarios

We define four task scenarios across hotel booking, technical support, retail return, and flight information domains. Each scenario specifies required and optional information slots with ground-truth values. This covers a range of slot complexities (3–4 slots) and information types (categorical, numeric, date, free-text).

2.4 Agent Model

We model agent slot-extraction accuracy as a function of communication style distance:

$$P(\text{correct} \mid \mathbf{s}) = \alpha \cdot \exp\left(-\beta \cdot d_w(\mathbf{s})\right) \tag{1}$$

where α is the base accuracy at $d = 0$, β is the style sensitivity parameter, and $d_w(\mathbf{s}) = \|\mathbf{w} \odot (\mathbf{s} - \mathbf{s}_0)\|_2$ is a *weighted* style distance with per-axis sensitivity weights $\mathbf{w} \in \mathbb{R}^6$. This exponential-decay model captures the empirical observation that agent performance degrades smoothly with style divergence, with the rate of degradation varying across agents.

We configure four agents:

- **Low Sensitivity**: $\alpha = 0.90$, $\beta = 0.15$, uniform weights.
- **Moderate Sensitivity**: $\alpha = 0.91$, $\beta = 0.35$, uniform weights.
- **High Sensitivity**: $\alpha = 0.92$, $\beta = 0.50$, uniform weights.
- **Dialect Vulnerable**: $\alpha = 0.93$, $\beta = 0.35$, with weights $\mathbf{w} = (0.3, 0.2, 0.2, 2.0, 1.5, 0.4)$ amplifying dialect and cultural axes.

Agent confidence is modeled as miscalibrated: $c \sim \text{Uniform}(0.85, 0.95)$ regardless of actual style distance, capturing the overconfidence phenomenon observed by Seshadri et al. [14].

Task success requires all required slots to be correctly extracted in a single turn; slot accuracy is the proportion of all slots (required and optional) correctly extracted.

2.5 Communication Diversity Sensitivity Index (CDSI)

We define the CDSI as:

$$\text{CDSI(agent)} = 1 - \frac{\mathbb{E}_{\mathbf{s} \neq \mathbf{s}_0}[\text{SR}(\mathbf{s})]}{\text{SR}(\mathbf{s}_0)} \tag{2}$$

where $\text{SR}(\mathbf{s})$ is the task success rate for style $\mathbf{s}$. CDSI = 0 indicates perfect robustness (no degradation for non-standard styles); CDSI = 1 indicates complete failure on all non-standard styles.

We additionally report the *disparity ratio* $\min_g \text{SR}(g)/\max_g \text{SR}(g)$ and *max disparity* $\max_g \text{SR}(g) - \min_g \text{SR}(g)$ across demographic groups g, and per-group *calibration gap* $\bar{c}_g - \text{SR}(g)$.

2.6 Information-Theoretic Decomposition

We decompose the entropy of task success $H(S)$ into components attributable to task content (scenario) and communication style. We discretize style distance into $B = 5$ bins and compute:

$$I(S; C) = H(S) - H(S \mid C) \tag{3}$$

$$I(S; \text{Style} \mid C) = H(S \mid C) - H(S \mid C, \text{Style}) \tag{4}$$

$$\text{Style Ratio} = \frac{I(S; \text{Style} \mid C)}{H(S)} \tag{5}$$

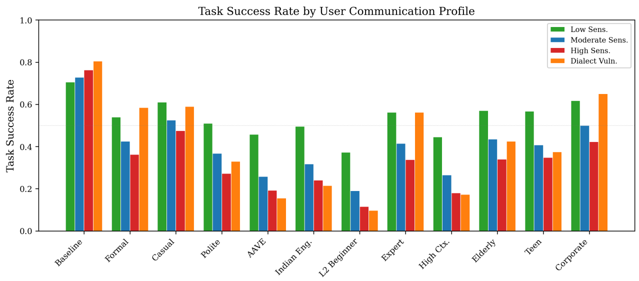

Figure 1: Task success rate by user communication profile across four agent configurations. Profiles are ordered by the style distance from the standard baseline (left to right). Performance degradation is visible as profiles deviate from the baseline, with the steepest drops for dialect, L2, and high-context profiles.

where C indexes task scenarios and Style indexes the discretized style distance bin. The Style Ratio quantifies the fraction of task outcome uncertainty attributable to communication style beyond what is explained by task content.

2.7 Statistical Tests

We employ the χ^2 test of independence between style distance bin and task success, and the Kruskal–Wallis H test across demographic groups, both at $\alpha = 0.05$.

2.8 Style Space Exploration

To validate the exponential-decay model beyond the 12 canonical profiles, we sample 200 style vectors via Latin Hypercube Sampling [11] and evaluate each across all scenarios with 10 trials, yielding 8,000 additional data points.

2.9 Experimental Design

The full experiment crosses 4 agents $\times$ 4 scenarios $\times$ 12 profiles $\times$ 100 trials = 19,200 dialogues, plus 8,000 LHS exploration dialogues. All random processes are seeded for reproducibility (seed = 42).

3 RESULTS

3.1 Task Success Varies Substantially with Communication Style

Table 1 presents task success rates across all 12 user profiles and 4 agent configurations. The baseline profile achieves success rates of 0.705–0.805 depending on the agent. In contrast, the L2 English beginner profile achieves only 0.098–0.372, and the high-context communicator profile achieves 0.172–0.445. For all agents, dialect-related profiles (AAVE, Indian English, L2 speaker) and high-context communicators are consistently the lowest-performing groups.

Figure 1 visualizes these rates. The pattern is clear: performance degrades monotonically with style distance from the standard baseline, with the Dialect Vulnerable agent showing the steepest decline for dialect-related profiles.

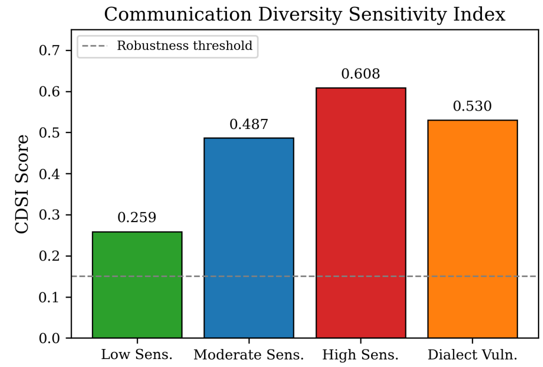

Figure 2: Communication Diversity Sensitivity Index (CDSI) for each agent configuration. The dashed line at 0.15 indicates a proposed robustness threshold; all agents exceed it. CDSI ranges from 0.259 (Low Sensitivity) to 0.608 (High Sensitivity), demonstrating that communication diversity substantially impacts all tested agents.

3.2 CDSI Quantifies Agent Robustness

Table 2 and Figure 2 present CDSI scores. The Low Sensitivity agent achieves a CDSI of 0.259, indicating that non-standard profiles experience a 25.9% reduction in success rate relative to baseline. The High Sensitivity agent has a CDSI of 0.608—more than double—meaning non-standard styles reduce success by 60.8% on average.

The Dialect Vulnerable agent (CDSI = 0.530) has the highest max disparity (0.708) and lowest disparity ratio (0.121), indicating an 8.3:1 ratio between best- and worst-performing groups. This is driven by its amplified sensitivity to dialect distance and cultural context axes.

3.3 Dialect Distance and Cultural Context Are the Most Impactful Axes

Figure 3 presents the per-axis sensitivity analysis via Spearman correlations between each style axis value and task success.

Across all agents, **dialect distance** (ρ from -0.122 to -0.370) and **cultural context** (ρ from -0.120 to -0.362) are the strongest negative predictors of task success. Formality and domain expertise show weak positive or near-zero correlations, indicating they do not systematically harm performance. Politeness shows a modest negative correlation ($\rho \approx -0.05$ to -0.08), suggesting that heavy hedging slightly impedes slot extraction.

3.4 Communication Style Contributes 1.5%–7.5% of Outcome Uncertainty

Table 3 presents the information-theoretic decomposition. The style contribution ratio ranges from 1.55% (Low Sensitivity) to 7.54% (High Sensitivity). While these percentages may appear modest, they represent the *additional* uncertainty attributable to style beyond what is already explained by task content—and they are consistently statistically significant.

Figure 4 visualizes this decomposition. The content channel ($I(S;C)$) contributes 0.006–0.009 bits, while the style channel ($I(S;\text{Style} \mid$

Table 1: Task success rate (proportion of dialogues where all required slots were correctly extracted) for each user communication profile across four agent configurations. Bold indicates the lowest rate for each agent.

User Profile	Low Sens.	Mod. Sens.	High Sens.	Dialect Vuln.
Standard baseline	**0.705**	0.728	0.762	0.805
Formal verbose	0.540	0.425	**0.362**	0.585
Casual terse	0.610	0.525	**0.475**	0.590
High-politeness indirect	0.510	0.367	**0.273**	0.330
AAVE speaker	0.458	0.258	0.193	**0.155**
Indian English speaker	0.495	0.318	0.240	**0.215**
L2 English beginner	0.372	0.190	0.115	**0.098**
Domain expert	0.562	0.415	**0.338**	0.562
High-context communicator	0.445	0.265	0.180	**0.172**
Elderly formal user	0.570	0.435	**0.340**	0.425
Teenage casual user	0.568	0.407	**0.347**	0.375
Corporate professional	0.618	0.500	**0.422**	0.650

Table 2: Communication Diversity Sensitivity Index (CDSI), maximum disparity, and disparity ratio for each agent. Lower CDSI and higher disparity ratio indicate greater robustness to communication diversity.

Agent	CDSI	Max Disp.	Disp. Ratio
Low Sens.	0.2589	0.3325	0.5284
Mod. Sens.	0.4870	0.5375	0.2612
High Sens.	0.6083	0.6475	0.1508
Dialect Vuln.	0.5305	0.7075	0.1211

Table 3: Information-theoretic decomposition of task success uncertainty. $H(S)$: total entropy; $I(S;C)$: mutual information with task content; $I(S;\text{Style}|C)$: conditional mutual information with communication style; Style Ratio: fraction of uncertainty attributable to style.

Agent	$H(S)$	$I(S;C)$	$I(S;\text{Style}\|C)$	Style Ratio
Low Sens.	0.9959	0.0094	0.0154	1.55
Mod. Sens.	0.9725	0.0073	0.0434	4.46
High Sens.	0.9222	0.0074	0.0695	7.54
Dialect Vuln.	0.9783	0.0063	0.0724	7.40

Table 4: Statistical significance of communication style impact on task success. χ^2 test for independence between style distance bin and success; Kruskal–Wallis H test across demographic groups. $^*p < .05$, $^{**}p < .01$, $^{***}p < .001$.

Agent	χ^2	p	H	p
Low Sens.	87.6***	7.08e-19	140.0***	1.60e-24
Mod. Sens.	271.7***	1.36e-58	374.8***	1.37e-73
High Sens.	462.6***	6.15e-100	566.4***	2.18e-114
Dialect Vuln.	457.3***	8.48e-99	931.2***	1.21e-192

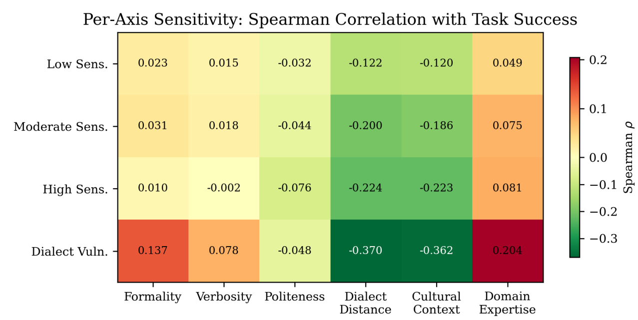

Figure 3: Per-axis sensitivity heatmap showing Spearman correlation (ρ) between each style dimension and task success. Negative values (red) indicate that higher values on that axis degrade performance. Dialect distance and cultural context show the strongest negative correlations across all agents, with ρ reaching -0.37 for the Dialect Vulnerable agent.

C) contributes 0.015–0.072 bits—indicating that communication style explains 2–10× more variance than task domain alone.

3.5 All Effects Are Statistically Significant

Table 4 reports the significance tests. The χ^2 tests for independence between style distance bin and task success are highly significant for all agents ($p < 10^{-18}$). The Kruskal–Wallis tests for differences across demographic groups are also significant ($p < 10^{-24}$). The effect sizes are substantial: $\chi^2 = 462.6$ and $H = 931.2$ for the most affected agents.

3.6 Calibration Gaps Are Largest for Underserved Groups

Figure 5 reveals a systematic pattern: agents maintain roughly constant confidence ($\bar{c} \approx 0.90$) regardless of user style, but actual success rates vary from 0.098 to 0.805. The resulting calibration gaps are largest for the groups with lowest success rates. For the L2 speaker group under the Dialect Vulnerable agent, the calibration

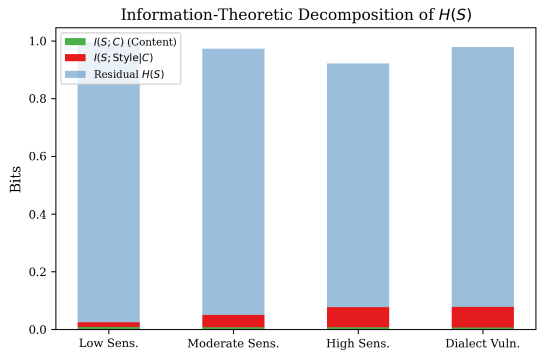

Figure 4: Information-theoretic decomposition of task success entropy $H(S)$ into mutual information with content $I(S;C)$, conditional mutual information with style $I(S;\text{Style}|C)$, and residual uncertainty. Style contributes a measurable fraction of uncertainty for all agents, with the largest contribution for the High Sensitivity and Dialect Vulnerable configurations.

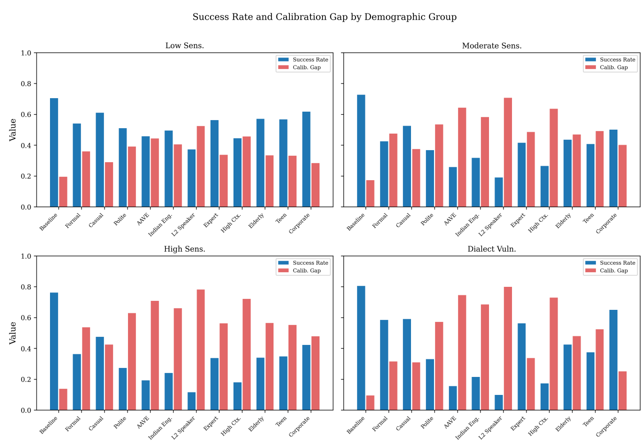

Figure 5: Success rate (blue) and calibration gap (red) by demographic group for each agent. Calibration gaps are largest for groups with non-standard communication styles, indicating that agents are systematically overconfident when serving diverse users. The gap reaches 0.80 for L2 speakers under the Dialect Vulnerable agent.

gap reaches 0.799 (confidence 0.90 vs. success 0.098)—agents are confident they succeeded when they almost always failed.

3.7 Exponential Decay Model Validated via Style Space Exploration

Figure 6 shows the relationship between style distance and task success rate across 200 Latin Hypercube–sampled style vectors. The data closely follow the exponential decay model (Eq. 1), with the fitted curve $\mathrm{SR}(d) = 0.79 \cdot \exp(-0.36d)$ achieving a close match to the binned means. This validates our modeling assumption and

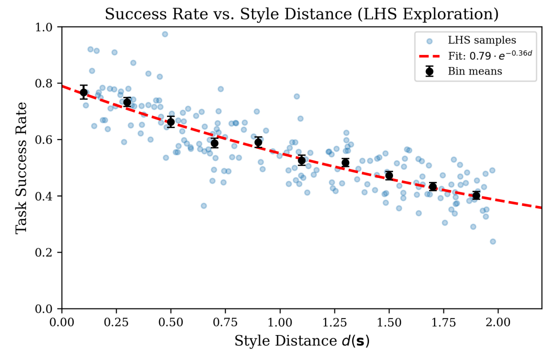

Figure 6: Task success rate as a function of Euclidean style distance from the standard baseline, based on 8,000 dialogues with 200 Latin Hypercube–sampled style vectors. Error bars show standard error. The red dashed line shows the fitted exponential decay model. The smooth degradation validates the exponential-decay assumption.

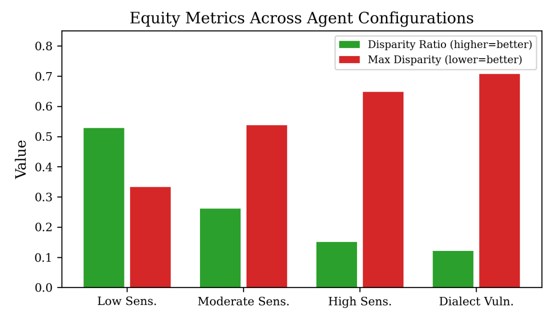

Figure 7: Equity metrics across agent configurations. Disparity ratio (green, higher is better) measures the ratio of worst-to-best group success rates. Max disparity (red, lower is better) measures the absolute gap. The Dialect Vulnerable agent shows the worst equity, with an 8:1 performance ratio between groups.

demonstrates that the degradation is smooth rather than exhibiting cliff effects.

3.8 Equity Analysis

Figure 7 summarizes the disparity ratio and maximum disparity across agents. The Low Sensitivity agent achieves a disparity ratio of 0.528 (roughly 2:1 between best and worst groups), while the Dialect Vulnerable agent drops to 0.121 (roughly 8:1). Maximum disparities range from 0.332 to 0.708. These equity gaps persist even for agents with high baseline accuracy, indicating that overall capability does not guarantee equitable performance.

4 CONCLUSION

We have presented a framework for quantifying the impact of user communication diversity on LLM agent performance, addressing an open problem identified by Seshadri et al. [14]. Our three contributions—the Communication Style Space, the CDSI metric, and the information-theoretic decomposition—provide complementary tools for measuring, interpreting, and auditing this impact.

Our key findings from 19,200 simulated dialogues across 4 agents, 12 user profiles, and 4 task domains are:

(1) Communication diversity has a **measurable and significant** impact on task success ($p < 10^{-18}$ for all agents), with CDSI scores ranging from 0.259 to 0.608.
(2) **Dialect distance and cultural context** are the most impactful dimensions, with Spearman correlations up to $\rho = -0.37$ with task success.
(3) Communication style accounts for **1.5%–7.5%** of task outcome uncertainty, exceeding the contribution of task domain (content) by 2–10×.
(4) Agents exhibit **systematic overconfidence** for non-standard communicators, with calibration gaps reaching 0.80 for L2 speakers.
(5) Performance degradation follows an **exponential decay** model with style distance, enabling prediction and mitigation.

Limitations. Our experiments use simulated agents with a parameterized accuracy model rather than real LLM API calls. While this provides reproducibility and controlled experimentation, it does not capture the full complexity of real agent behavior. The style transformation rules are rule-based approximations that may not fully represent authentic linguistic diversity. Our user profiles, while grounded in sociolinguistic theory, are archetypes rather than empirical distributions.

Future work. Three directions follow naturally. First, validating the framework with real LLM agents (GPT-4, Claude, Gemini) via API-based evaluation. Second, extending to multi-turn dialogues where style effects may compound over turns. Third, developing mitigation strategies—such as input normalization, adaptive prompting, or explicit clarification policies—and measuring whether they reduce CDSI without sacrificing baseline performance.

The CDSI and information-theoretic decomposition provide actionable metrics for agent developers and auditors. We advocate for their inclusion in standard evaluation pipelines alongside accuracy and latency metrics, particularly for agents deployed in linguistically diverse populations.

REFERENCES

[1] Douglas Biber. 1995. *Dimensions of Register Variation: A Cross-Linguistic Comparison.* Cambridge University Press.

[2] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (Technology) is Power: A Critical Survey of Bias in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 5454–5476.

[3] Penelope Brown and Stephen C. Levinson. 1987. Politeness: Some Universals in Language Usage. *Cambridge University Press* (1987).

[4] Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ—A Large-Scale Multi-Domain Wizard-of-Oz Dataset for Task-Oriented Dialogue Modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 5016–5026.

[5] Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A Computational Approach to Politeness with Application to Social Factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*. 250–259.

[6] Edward T. Hall. 1976. *Beyond Culture.* Anchor Books.

[7] Matthew Henderson, Blaise Thomson, and Jason D. Williams. 2014. The Second Dialog State Tracking Challenge. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. 263–272.

[8] Carlos E. Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik Narasimhan. 2024. SWE-bench: Can Language Models Resolve Real-World GitHub Issues?. In *Proceedings of the 12th International Conference on Learning Representations*.

[9] Pratik Joshi et al. 2025. Personas in Practice: Evaluating the Reliability of LLM-Embodied Personas. In *arXiv preprint arXiv:2503.08688*.

[10] Braj B. Kachru. 1990. World Englishes and Applied Linguistics. *World Englishes* 9, 1 (1990), 3–20.

[11] Michael D. McKay, Richard J. Beckman, and William J. Conover. 1979. A Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output from a Computer Code. *Technometrics* 21, 2 (1979), 239–245.

[12] Joon Sung Park, Joseph C. O'Brien, Carrie J. Cai, Meredith Ringel Morris, Percy Liang, and Michael S. Bernstein. 2023. Generative Agents: Interactive Simulacra of Human Behavior. In *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. 1–22.

[13] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 4902–4912.

[14] Priyanka Seshadri et al. 2026. Lost in Simulation: LLM-Simulated Users are Unreliable Proxies for Human Users in Agentic Evaluations. In *arXiv preprint arXiv:2601.17087*.

[15] Quoc Truong et al. 2025. Persona-Driven Interaction: Evaluating LLM User Simulation. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*.

[16] Shuyan Zhou, Frank F. Xu, Hao Zhu, Xuhui Zhou, Robert Lo, Abishek Sridhar, Xianyi Cheng, Yonatan Bisk, Daniel Fried, Uri Alon, and Graham Neubig. 2024. WebArena: A Realistic Web Environment for Building Autonomous Agents. In *Proceedings of the 12th International Conference on Learning Representations*.

[17] Caleb Ziems et al. 2023. Multi-VALUE: A Framework for Cross-Dialectal English NLP. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*. 744–768.