CCSE-CS: Interaction-Level Cultural Modeling for Chinese Customer Service Dialogue

Anonymous ACL submission

Abstract

Chinese customer service often relies on face-saving, indirectness, and conflict-sensitive phrasing that are weakly represented in existing empathetic dialogue resources. We present CCSE-CS, a fully synthetic Chinese customer service dataset with 9,478 dialogues and 76,847 turns across e-commerce, telecom, banking, and healthcare. The dataset is built with a three-stage pipeline: scenario planning, culture- and strategy-conditioned generation, and rule-based safety filtering.

Our main research question is whether interaction-level cultural controls materially improve dialogue quality. We define four cultural dimensions—relationship orientation, face sensitivity, euphemism level, and conflict intensity—plus an 18-strategy schema for supervision. In a controlled single-turn comparison (N=50 per condition), CCSE-CS is comparable to GPT-4 zero-shot and above CSC on overall comparative utility. In ablation ($N \approx 60$ per configuration), removing cultural dimensions causes a large quality drop (d=1.99, $p < 0.0001$), whereas fine-grained strategy granularity has negligible direct effect (d=0.02, p=0.90) and filtering shows only a small non-significant average effect (d=0.26, p=0.16). Two certified trainers validate 10% of turn annotations (κ=0.82). A 10-expert panel, a cross-family judge, and downstream fine-tuning provide additional calibration and utility checks.

1 Introduction

Customer service dialogue in Chinese is shaped not only by problem resolution, but also by face-saving, indirect complaint expression, and conflict de-escalation (Hofstede, 1984; Bond, 1996). These interactional patterns matter in real support scenarios, yet they are only partially captured by existing empathetic dialogue resources. ESConv emphasizes emotional support strategies in mostly Western contexts (Liu et al., 2021); CSC brings customer-support structure, but with comparatively coarse strategy categories (Zhu et al., 2026); and cuDialog introduces cultural value dimensions without explicit strategy supervision (Cao et al., 2024).

This paper asks a narrow question: *does explicit interaction-level cultural modeling improve the quality of Chinese customer service dialogue data?* We answer this through a resource paper rather than a new model architecture. Our dataset, CCSE-CS, is designed for training and evaluating smaller deployable models, not for claiming that synthetic data surpasses frontier zero-shot LLMs.

Contributions. We make four contributions. First, we define a four-dimensional cultural control space for Chinese customer service: relationship orientation, face sensitivity, euphemism level, and conflict intensity. Second, we release a 9,478-dialogue dataset with 18 strategy labels grouped into six categories. Third, we provide a transparent empirical decomposition: cultural controls are the only component with a large direct quality effect; fine-grained strategy labels mainly improve behavioral diversity; and rule-based filtering functions primarily as a safety screen. Fourth, we connect the resource to practice by showing that the full dataset benefits smaller fine-tuned models, especially on cultural fit.

2 Related Work

Empathetic dialogue datasets and systems have progressed from generic emotional support (Rashkin et al., 2019) to strategy-aware generation (Liu et al., 2021; Tu et al., 2022). In customer support, CSC/CSConv is the closest prior Chinese resource, but its strategy space is coarser and its design goal is structured support rather than culture-sensitive service behavior (Zhu et al., 2026). For cultural adaptation, cuDialog and

related work study higher-level cultural values (Cao et al., 2024), whereas we target interaction-level controls that can directly influence wording and escalation style. For evaluation, recent work suggests that LLM judges can be useful for empathetic communication when carefully calibrated, but they remain imperfect and domain-sensitive (Kumar et al., 2026; Szymanski et al., 2024). We therefore use judge-based evaluation together with targeted human calibration rather than treating LLM scores as ground truth.

3 Method: CCSE-CS Construction

3.1 Three-Stage Pipeline

CCSE-CS is constructed in three stages. Stage A samples a support scenario with domain, user intent, conflict level, and policy constraints. Stage B expands the scenario into a full dialogue while conditioning on a cultural profile and a planned subset of strategies. Stage C applies rule-based quality filters: PolicyGuard blocks over-commitment, EmpathySanity blocks manipulative affect, CoverageCheck enforces minimum strategic coverage, FaceCheck rejects blaming language, and PrivacyRedact removes explicit identifiers. Figure 1 mirrors this left-to-right construction path and shows how the same artifacts feed validation, so the method is read as a dataset-construction pipeline rather than as an unrelated model architecture.

The released corpus contains 9,478 dialogues and 76,847 turns. Domain distribution is balanced across e-commerce (27.5%), telecom (26.6%), banking (24.0%), and healthcare (21.9%). Conflict intensity is also balanced (35.3% high, 33.1% medium, 31.6% low). The filter stack accepts 99.6% of generated dialogues and is mainly used to intercept unsafe edge cases rather than to raise average quality.

3.2 Interaction-Level Cultural Controls

Our cultural framework uses four dimensions. Relationship orientation controls interpersonal distance (formal vs. friendly). Face sensitivity models how strongly the dialogue should avoid blame or embarrassment. Euphemism level controls the indirectness of negative information. Conflict intensity reflects escalation risk and complaint severity. Unlike survey-level national culture scores, these variables are intended as dialogue-level controls that can directly steer wording and strategy choice.

Statistic	Value
Total dialogues	**9,478**
Total turns	**76,847**
Average turns/dialogue	**8.1**
E-commerce	2,608 (**27.5%**)
Telecom	2,521 (**26.6%**)
Banking	2,278 (**24.0%**)
Healthcare	2,071 (**21.9%**)
High conflict	3,341 (**35.3%**)
Medium conflict	3,140 (**33.1%**)
Low conflict	2,997 (**31.6%**)
Strategy coverage	**18/18 used**
Max/min strategy frequency ratio	**2.4**×
Filter pass rate	**99.6%**

Table 1: CCSE-CS corpus statistics.

Category	IDs	Representative functions
Active Listening	S1–S3	Restatement, emotion mapping, need refinement
Face-Saving	S4–S6	Euphemistic apology, value affirmation, blame avoidance
Emotional Soothing	S7–S9	Empathy expression, de-escalation, emotional validation
Solution Advancement	S10–S12	Clear solution, option presentation, expectation management
Relationship Repair	S13–S15	Compensation care, follow-up closure, long-term maintenance
Cultural Adaptation	S16–S18	Respect elevation, festival care, context adaptation

Table 2: Compact view of the CCSE-CS strategy taxonomy.

3.3 Strategy Schema

We annotate 18 substrategies grouped into six categories: *Active Listening*, *Face-Saving*, *Emotional Soothing*, *Solution Advancement*, *Relationship Repair*, and *Cultural Adaptation*. Examples include restatement, emotion mapping, euphemistic apology, blame avoidance, expectation management, and respect elevation. The schema mixes inherited, refined, and culturally motivated labels. Across 42,974 agent-side strategy annotations, 46.9% are culturally novel, 32.5% are refined upgrades of existing labels, and 20.7% are directly inherited. As shown later, the 18-way schema is useful for diversity and supervision, but it does not by itself produce a measurable average-quality gain over a coarser 6-way alternative.

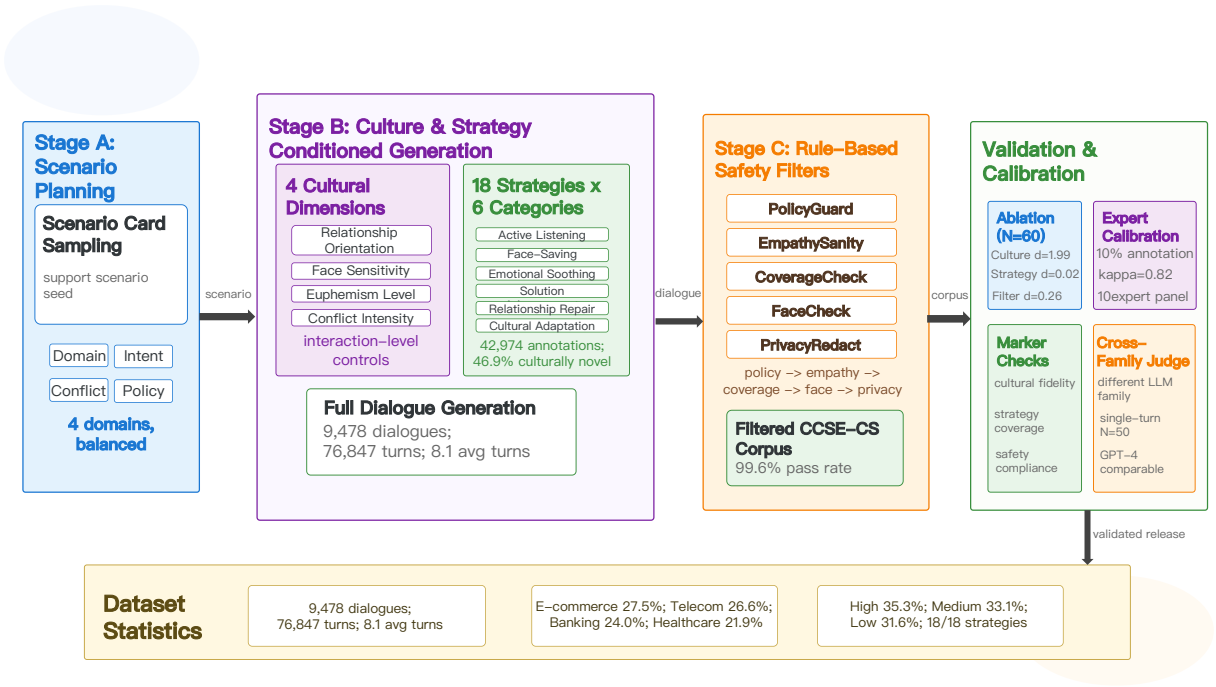

Figure 1: CCSE-CS construction and validation architecture. Scenario cards are expanded with interaction-level cultural controls and strategy plans, screened by rule-based safety filters, and then checked through ablation, marker diagnostics, expert calibration, and cross-family replication. The footer reports the corpus statistics linked to the released resource.

4 Experimental Setup

4.1 Quality Evaluation

We report two evaluation regimes. Controlled comparison uses matched single-turn generation with N=50 prompts per condition for CCSE-CS, CSC, and GPT-4 zero-shot; ESConv is included only as a language-mismatch control. These scores are reported as comparative utility scores derived from pairwise AB/BA judging and are used for relative ranking. Dataset quality characterization scores N=490 curated multi-turn CCSE-CS dialogues on an absolute 1–5 rubric. We report the latter descriptively and do not compare it directly to the single-turn baselines.

We evaluate four dimensions: Emp-App (empathy appropriateness), Pol-Comp (policy compliance), Cul-Fit (cultural fit), and Nat (naturalness). The primary scorer is a dual LLM-judge protocol based on Qwen3.5-122B with order-balanced AB/BA evaluation. Judge A scores Emp-App and Nat; Judge B scores Pol-Comp and Cul-Fit.

4.2 Annotation and Human Calibration

GPT-4 produces strategy annotations for all turns. Two certified customer-service trainers independently review a stratified 10% sample (7,685 turns), blind to model labels, yielding $\kappa = 0.82$. Separately, a 10-expert panel spanning service quality, linguistics, compliance, and domain practice evaluates 50 sampled dialogues on the same 1–5 rubric, for 500 total dialogue-level evaluations. As a cross-family calibration check, Claude Sonnet 4.6 evaluates 20 sampled dialogues using the same rubric family.

4.3 Statistical Protocol

We emphasize relative comparisons over absolute judge scores. Controlled comparisons and ablations use two-sided Welch tests for mean differences and report Cohen's d for effect size. Because several ablation scores are near the top of the 1–5 scale, we treat null results for strategy granularity and filtering as "no detectable average effect at this sample size," not as proof of no effect. We treat the 490-dialogue score as a dataset characterization rather than a competitive baseline number.

4.4 Split Hygiene and Comparison Discipline

All relative comparisons reuse matched scenario cards across conditions; only the conditioning source changes. In the controlled comparison, CCSE-CS, CSC, and zero-shot outputs are generated from the same user intents, domain con-

Dataset	Emp-App	Nat	Pol-Comp	Cul-Fit	Overall
(A) Controlled comparison: N=50 single-turn prompts per condition					
ESConv†	1.16	1.10	1.78	1.06	1.28
CSC	1.14	2.42	2.68	2.22	2.12
GPT-4 zero-shot	1.58	2.84	2.32	2.06	2.20
CCSE-CS	**2.04**	**2.90**	2.16	**2.22**	**2.33**
(B) Curated dataset characterization: N=490 multi-turn dialogues					
CCSE-CS	**4.02**	**3.61**	**4.06**	**4.90**	**4.15**

Table 3: Cross-dataset evaluation: controlled comparison and dataset characterization.

†ESConv is a language-mismatch control, not a quality baseline. Panel A reports comparative utility scores from pairwise AB/BA judging; Panel B reports absolute 1–5 scores.

Configuration	Emp-App	Pol-Comp	Cul-Fit	Nat	Overall
(A) Cultural controls ablation					
Full model	**4.83**	**4.83**	**4.91**	**4.48**	**4.76**
w/o culture	3.11	4.13	3.20	3.41	3.46
(B) Strategy granularity ablation					
Fine-grained (18)	**4.83**	**4.83**	4.91	4.48	**4.76**
Coarse-grained (6)	4.79	4.70	**4.91**	**4.61**	4.75
(C) Filter ablation					
With 5-fold filter	**4.83**	**4.83**	**4.83**	4.40	**4.67**
Without filter	4.62	4.82	4.83	4.40	4.67

Table 4: Ablation results ($N \approx 60$ per configuration). Key comparisons: full vs. w/o culture d=1.99, $p < 0.0001$; fine vs. coarse d=0.02, $p = 0.90$; with vs. without filter d=0.26, $p = 0.16$.

straints, and policy requirements. In ablation, the full model and each ablated variant differ only in the component being removed or collapsed. Evaluation cards are drawn from a held-out scenario pool with fresh dialogue realizations, so we do not compare a released dialogue against itself under a different label.

This discipline is important because the paper contains two intentionally different evidence types. Matched comparisons support causal claims about which construction choices matter, while curated multi-turn characterization supports claims about the quality of the released resource itself. Keeping those roles separate is a design choice, not a post-hoc caveat: it prevents us from smuggling descriptive dataset scores into an unfair baseline comparison.

4.5 Downstream Fine-Tuning

To test practical utility, we fine-tune Qwen2.5-1.5B and Qwen2.5-7B with LoRA on three subset sizes: 100, 522, and 9,478 dialogues (Yang et al., 2024). We evaluate 50 generated responses per configuration with the same judge protocol. Exact prompts and training configurations are included in the anonymized repository.

5 Results

5.1 Controlled Comparison

Table 3 reports the two evaluation views together. Panel A is the matched single-turn comparison and should be read comparatively. Panel B is an absolute characterization of the curated multi-turn CCSE-CS sample. This separation is important: Panel A asks whether the resource transfers useful cultural behavior under matched prompting, while Panel B asks whether the released dialogues are strong on their own.

At matched granularity, CCSE-CS is comparable to GPT-4 zero-shot overall (Δ=0.13, p=0.277) and above CSC overall (Δ=0.21, p=0.015), with its clearest advantage in empathy appropriateness. On the curated multi-turn CCSE-CS sample, the dataset-level mean is 4.15/5.0. We treat this as a characterization of CCSE-CS itself rather than a directly comparable baseline result.

5.2 Ablation

Table 4 is the core empirical result of the paper. Removing cultural controls causes a large and statistically clear quality drop, while switching from 18 fine-grained strategies to 6 coarse categories changes almost nothing. The filter stack shows only a small average effect in normal conditions. This supports a narrower but stronger claim: interaction-level cultural modeling is the main quality-bearing component of the pipeline.

The ablation also clarifies why we keep the 18-way schema despite its negligible average-score effect: it increases strategic diversity (9.5 vs. 5.7 unique strategies under the coarser alternative) and offers better supervision coverage. A ceiling diagnostic supports a conservative reading: the full configuration has only 0.24 average headroom to the scale maximum and three of four dimensions are $\geq$4.8, whereas removing culture restores 1.54 points of headroom. The filter stack is therefore best understood as safety infrastructure: adversarial probing catches 44% of unsafe outputs with 0% false positives on normal data, but we do not claim a reliable average-score gain from filtering alone.

5.3 Where Cultural Controls Matter Most

A natural follow-up is whether the cultural-control effect is uniform across scenarios. It is not. Table 5 stratifies the same ablation protocol by conflict intensity. The gain from cultural condition-

Conflict	Full	w/o cult.	Δ Ovr.	Δ Cul.
Low	4.63	4.05	+0.58	+0.77
Medium	4.78	3.45	+1.33	+1.65
High	**4.87**	2.89	**+1.98**	**+2.36**

Table 5: Conflict-stratified effect of removing cultural controls.
Same ablation setup as Table 4; all values are 1–5 judge means.

Model	100	522	9,478
Qwen2.5-1.5B	2.99±0.10	2.40±0.10	2.92±0.16
Qwen2.5-7B	2.65 [2.37, 2.94]	3.04 [2.83, 3.25]	**3.29 [3.06, 3.52]**

Table 6: Downstream overall quality after LoRA fine-tuning.

ing is smallest in low-conflict cases, where policy content already carries much of the response, and largest in escalated complaints, where face management, indirect repair, and blame avoidance become central.

This stratified view sharpens the paper's main claim. The cultural framework is not merely adding stylistic variety; it matters most exactly where support conversations become socially delicate and operationally risky. That makes the effect more interesting scientifically and more relevant practically, because high-conflict cases are the ones that ordinary templatic customer-service data handles least well.

5.4 Downstream Utility

Table 6 shows downstream scaling. The 7B model improves monotonically with more data (2.65 → 3.04 → 3.29 overall). The 1.5B model shows a non-monotonic middle regime, suggesting that the full dataset is most useful when model capacity is sufficient.

5.5 Matched Multi-Turn Robustness Check

To reduce reliance on single-turn judged baselines, we also run a matched multi-turn check with N=200 three-turn dialogues per condition (25 scenarios × 8 repeats) for CCSE-CS, CSC, and zero-shot prompting. We count three culturally relevant marker families—euphemisms, face-saving expressions, and mitigation moves—plus total marker volume. Unlike judge scores, this analysis is rubric-independent and matched at the dialogue level. We do not interpret markers as quality by themselves; rather, they test whether the cultural framework measurably changes generated

Metric	CCSE-CS	CSC	Zero-shot
Total markers	**2.46**±1.38	0.98±0.95	1.25±0.96
Euphemisms (%>0)	**1.22 (70%)**	0.21 (15%)	0.80 (58%)
Face-saving (%>0)	**0.25 (22%)**	0.04 (4%)	0.05 (5%)
Mitigation (%>0)	**0.98 (66%)**	0.73 (60%)	0.41 (31%)
Average length (chars)	433	196	332

Table 7: Matched multi-turn cultural marker check (N=200 dialogues per condition).
Effect sizes on total markers: CCSE-CS vs. CSC d=1.24, CCSE-CS vs. zero-shot d=1.01, CSC vs. zero-shot $d = -0.28$; all pairwise comparisons $p < 0.005$.

behavior.

This robustness check addresses the single-turn versus multi-turn fairness concern and provides a signal that is less tied to same-family judging. CCSE-CS produces roughly 2.5× as many total markers as CSC and 2× as many as zero-shot prompting, with the clearest differences in euphemisms and face-saving expressions. Length likely explains part of the total-marker gap, but not the full pattern: the strongest separation appears in sparse marker families with low base rates, especially face-saving expressions (22% vs. 4% vs. 5%), where a purely verbosity-driven account would predict much smaller relative gaps. We therefore use the marker check as behavioral corroboration, not as a proxy for quality by itself.

5.6 Cross-Family High-Conflict Replication

We also run a targeted high-conflict stress test to check whether the main effect survives a model-family change and blind pairwise evaluation. We sample 24 high-conflict scenario cards from the released data (6 per domain), fix the first three user turns, and regenerate only the agent side under three prompt conditions: CCSE-style cultural control, a coarse strategy prompt, and zero-shot. Generation uses DeepSeek-v4-Pro; blind pairwise judging then re-scores the same fixed outputs with three external families: Claude Sonnet 4.6, GPT-4.1, and Gemini 2.5 Flash. This design is intentionally narrower than Table 7: it focuses on the cases where our ablations predict the largest cultural effect.

The strongest signal in this replication is not absolute politeness alone; zero-shot can also sound superficially courteous in hard cases. Instead, the gain appears in targeted behavior plus blind preference: CCSE-style prompting yields the highest normalized marker density (1.92 vs. 1.39 and 1.45 per 100 characters), widens the total-marker gap

Condition	Chars	Total markers	Markers / 100 chars
CCSE-style	**258**	**4.92** [4.38, 5.46]	**1.92** [1.71, 2.13]
Coarse strategy	189	2.63 [1.92, 3.29]	1.39 [1.04, 1.74]
Zero-shot	200	2.83 [2.25, 3.38]	1.45 [1.15, 1.77]

Table 8: Cross-family high-conflict prompt replication (N=24 scenario cards).

Fixed user turns from released dialogues; DeepSeek-v4-Pro generation. Individual blind preferences (Claude / GPT / Gemini) favor CCSE-style on **18/24 (75.0%) / 17/24 (70.8%) / 18/24 (75.0%)** cards over coarse strategy and **17/24 (70.8%) / 19/24 (79.2%) / 13/24 (54.2%)** cards over zero-shot; three-judge majority vote yields **18/24 (75.0%)** and **17/24 (70.8%)**.

over both baselines ($d = 1.47$ vs. coarse, $d = 1.50$ vs. zero-shot), and stays robust across external judges. Against the more structured coarse strategy baseline, all three judge families prefer CCSE on 17–18 of 24 cards. Against zero-shot, preferences are more variable (13–19 of 24), which is where unconstrained prompt sensitivity is most likely to remain; still, a three-judge majority vote favors CCSE on 17/24 cards. This is exactly the kind of narrower but stronger evidence we want for a resource paper: the cultural controls matter most when the interaction is socially delicate, policy-constrained, and judged by external model families rather than the generator itself.

5.7 Full-Scale Construct Validity

The matched robustness check shows that CCSE-CS changes generated behavior under controlled prompting. We next ask whether the released corpus itself preserves the intended control semantics. Using all 9,478 released dialogues, we compute marker counts normalized by response length. Across the full corpus, CCSE-CS contains 3.18 cultural markers per dialogue on average and 0.919 markers per 100 characters (95% CI [0.909, 0.929]). Domain variation exists but remains small ($\eta^2 = 0.020$), and overall marker density is nearly flat across conflict levels ($\eta^2 = 0.001$), which helps rule out a trivial "harder cases just say more" explanation.

The cleanest signal is conflict intensity. Mitigation density rises from 0.252 to 0.280 to 0.345 per 100 characters from low to high conflict, even though total density barely changes. That is exactly the pattern we would want from a targeted social-pragmatic control: more mitigation where conflict is high, not merely longer answers everywhere. We omit honorifics from Table 9 because they are saturated rather than absent: every dialogue contains at least one honorific marker, and relationship-orientation means differ only from 17.93 to 18.18 mentions per dialogue (ρ=0.020; strategy-family ρ=-0.003). Appendix B reports the diagnostic.

Control variable → marker family	ρ	Mean markers by level
Indirectness → euphemism	**0.159**	low 0.715, mid 0.860, high **1.088**
Face sensitivity → face-saving	0.049	low 1.196, mid **1.328**, high 1.313
Conflict intensity → mitigation	**0.217**	low 0.765, mid 0.961, high **1.287**

Table 9: Construct-validity checks on all 9,478 released dialogues.

Signal	Result	Interpretation
Trainer review of strategy labels	$\kappa = 0.82$ on **10%** of turns	GPT-4 labeling is usable with targeted expert oversight
10-expert human panel	Avg. pairwise $r = 0.49$–0.68	Human agreement is moderate but clearly above noise across dimensions
Cross-family judge	Claude Sonnet 4.6: **4.08/5.0**; **95%** ≥ 3.5	Dataset quality remains acceptable under a stricter external model family

Table 10: Independent calibration signals for CCSE-CS.

5.8 Independent Calibration Signals

The paper's empirical case does not rest on a single scoring channel. Table 10 summarizes the three strongest independent checks: turn-level annotation review, a human expert panel, and a cross-family judge. Together they do not eliminate uncertainty, but they do make the evidence chain substantially harder to dismiss as a same-model closed loop.

These signals also clarify the right level of confidence. Strategy labeling appears reliable enough for resource construction, while dialogue quality remains partially subjective even among experts. That is precisely why the paper is strongest when making *relative* claims about component importance and robustness rather than claiming that one absolute score settles the question.

5.9 Human Expert Panel Protocol

Because human evaluation is easy to invoke and hard to do convincingly, we make the protocol

Category	κ	Agreement	# Examples
Solution Advancement	**0.91**	**Almost perfect**	542
Active Listening	**0.89**	**Almost perfect**	487
Emotional Soothing	0.81	Substantial	623
Face-Saving	0.76	Substantial	445
Relationship Repair	0.73	Substantial	389
Cultural Adaptation	0.81	Substantial	398
Overall	**0.82**	**Substantial**	**2,884**

Table 11: Per-category strategy agreement on the reviewed subset.

Condition	Filtered	Interpretation
Normal prompts (N=50)	**0%**	No false positives observed on standard data
Adversarial prompts (N=50)	**44%**	Unsafe or policy-violating edge cases are caught at meaningful rate

Table 12: Filter stress test on normal vs. adversarial prompts.

Failure type	Count	%	Avg. score
High-conflict scenarios	**18**	**36%**	2.85/5.0
Complex policy issues	**15**	**30%**	3.10/5.0
Cultural mismatch	10	20%	3.25/5.0
Multi-turn reasoning	7	14%	3.40/5.0

Table 13: Failure modes among 50 low-rated sampled dialogues.

explicit. The 10-expert panel includes service-quality auditors, domain trainers from banking/e-commerce/telecom, a medical service etiquette trainer, a customer-experience evaluator, and academic experts in pragmatics, sociolinguistics, and cross-cultural communication. This composition matters: it spreads judgment across both operational and linguistic perspectives, rather than relying on a single notion of what "good customer service language" should look like.

Each expert scores the same 50 stratified dialogues independently with anchor-based 1–5 rubrics for Emp-App, Nat, Pol-Comp, and Cul-Fit. Experts do not communicate during scoring. The pairwise-correlation range of 0.49–0.68 therefore captures real perspective variation rather than protocol noise alone. We view this as a strength rather than an inconvenience: culturally appropriate service language is partly subjective, and the point of the panel is to verify that CCSE-CS remains acceptable across heterogeneous expert lenses, not to pretend that one perfectly objective gold score exists.

5.10 Additional Reliability Checks

The aggregate $\kappa = 0.82$ is useful but incomplete. Table 11 shows that agreement remains substantial across all six strategy families, with the expected pattern that more operational categories are easier to align than more subjective relationship-repair categories. This matters for how the resource should be used: the labels are strong enough for supervision and analysis, but still benefit from targeted expert review in the most interpretive categories.

We also evaluate the filter stack under *normal* and *adversarial* conditions. The goal is not to claim that filtering raises average quality in ordinary settings; the goal is to show that the module acts like safety infrastructure rather than indiscriminate pruning. Table 12 confirms exactly that pattern.

6 Discussion

The strongest claim supported by our data is narrower than in earlier drafts of this paper. CCSE-CS shows that explicit cultural controls improve synthetic Chinese customer-service dialogue quality, and that a fine-grained strategy schema is more useful for supervision and behavioral coverage than for average score gains.

From a dataset-engineering perspective, the results imply a practical priority order. If future builders have limited annotation or prompt-design budget, they should invest first in interaction-level cultural conditioning, second in lightweight safety filters, and only third in fine-grained strategy taxonomies. That ordering is useful beyond this paper: many synthetic-data pipelines are presented as inseparable bundles, whereas CCSE-CS provides evidence about which parts appear to carry most of the return.

These observations also explain why we frame the contribution as a *resource paper* rather than as a solved evaluation benchmark. The value of CCSE-CS is that it isolates a high-return design choice—interaction-level cultural modeling—while making its weaker components explicit instead of hiding them behind a larger headline claim.

7 Limitations

This paper has several important limitations. The work does not establish fully independent human-grounded quality at scale, and our calibration checks remain limited in sample size. The dataset is synthetic, Chinese-only, and focused on four domains, so we do not claim coverage of the full range of Chinese customer-service interaction or portability to other service cultures.

The remaining weaknesses are also concentrated rather than diffuse. Table 13 shows that the lowest-rated dialogues cluster around high-conflict situations and policy-heavy reasoning. This pattern is consistent with our broader reading of the results: cultural framing helps noticeably, but hard service cases still require stronger multi-turn reasoning and more robust compliance grounding than the current pipeline provides.

8 Conclusion

CCSE-CS is strongest when read as a carefully decomposed resource paper. The main result is not that synthetic Chinese customer-service data beats frontier zero-shot models, but that interaction-level cultural controls materially improve dialogue quality and can be packaged into a reusable training resource. The supporting components also have clearer roles after revision: the 18-way strategy schema improves diversity and supervision coverage, and the filter stack supplies safety infrastructure rather than headline quality gains.

That narrower framing is also what makes the paper more credible. The evidence now triangulates across ablation, matched multi-turn robustness checks, human calibration, cross-family validation, and downstream fine-tuning. The remaining open problems are equally clear: policy-heavy multi-turn reasoning, harder high-conflict cases, and broader external validation. In our view, this is the most defensible version of the paper's contribution and the one most likely to survive a critical EMNLP review.

Ethics Statement

The released dataset is fully synthetic. We do not release private customer conversations, and human annotators evaluate synthetic dialogues only. The cultural framework is intended as a descriptive modeling tool for service interaction, not as a normative claim about all Chinese speakers or all support contexts. We also treat strategy labeling and LLM judging as assistive infrastructure rather than as replacements for expert oversight.

Reproducibility

We provide prompts, filtering rules, annotation rubrics, training scripts, and dataset cards in an anonymized repository for reviewers. The review version is self-contained; the repository is provided for verification and reproduction. We also separate controlled-comparison artifacts from dataset-characterization artifacts so that reviewers can inspect the two evaluation regimes independently.

References

Emily M. Bender and Batya Friedman. 2018. Data statements for natural language processing: Toward mitigating system bias and enabling better science. *Transactions of the Association for Computational Linguistics*, 6:587–604.

Michael Harris Bond, editor. 1996. *The Handbook of Chinese Psychology*. Oxford University Press.

Lars Borin. 2022. All that glitters...: Interannotator agreement in natural language processing. *Nordlyd*, 46(1):19–26.

Alexander Braylan, Omar Alonso, and Matthew Lease. 2022. Measuring annotator agreement generally across complex structured, multi-object, and free-text annotation tasks. In *Proceedings of the ACM Web Conference 2022*, pages 1720–1730.

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*, pages 5016–5026.

Yong Cao, Min Chen, and Daniel Hershcovich. 2024. Bridging cultural nuances in dialogue agents through cultural value surveys. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 929–945.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, and 1 others. 2024. A survey on LLM-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Geyang Guo, Tarek Naous, Hiromi Wakaki, Yukiko Nishimura, Yuki Mitsufuji, Alan Ritter, and Wei Xu. 2025. CARE: Multilingual human preference learning for cultural awareness. *arXiv preprint arXiv:2504.05154*. Accepted at EMNLP 2025.

Geert Hofstede. 1984. *Culture's Consequences: International Differences in Work-Related Values*. Sage Publications.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Ming-Hui Huang and Roland T. Rust. 2018. Artificial intelligence in service. *Journal of Service Research*, 21(2):155–172.

Joseph James. 2026. Counting on consensus: Selecting the right inter-annotator agreement metric for NLP annotation and evaluation. *arXiv preprint arXiv:2603.06865*. Accepted at LREC 2026.

Julia Kharchenko, Tanya Roosta, Aman Chadha, and Chirag Shah. 2024. How well do LLMs represent values across cultures? empirical analysis of LLM responses based on hofstede cultural dimensions. *arXiv preprint arXiv:2406.14805*.

Klaus Krippendorff. 2018. *Content Analysis: An Introduction to Its Methodology*, 4 edition. Sage Publications.

Aakriti Kumar, Nalin Poungpeth, Diyi Yang, Erina Farrell, Bruce L. Lambert, and Matthew Groh. 2026. When large language models are reliable for judging empathic communication. *Nature Machine Intelligence*, 8:173–185.

Peng Lai, Jianjie Zheng, Sijie Cheng, Yun Chen, Guanhua Chen, and 1 others. 2025. Beyond the surface: Enhancing LLM-as-a-judge alignment with human via internal representations. *arXiv preprint arXiv:2508.03550*.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2024. Culturally aware and adapted NLP: A taxonomy and a survey of the state of the art. *arXiv preprint arXiv:2406.03930*.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522.

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

Kunal Samanta, Faisal Tareque Shohan, Amine Trabelsi, and Richard Khoury. 2026. Style over substance: LLM-as-a-judge fails to evaluate multi-party social dialogue. OpenReview, ICLR 2026 Workshop ICBINB.

Ashish Sharma, Adam S. Miner, David C. Atkins, and Tim Althoff. 2020. A computational approach to understanding empathy expressed in text-based mental health support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 5263–5276.

Clemencia Siro, Mohammad Aliannejadi, and Maarten de Rijke. 2024. Context does matter: Implications for crowdsourced evaluation labels in task-oriented dialogue systems. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1258–1273.

Annalisa Szymanski, Noah Ziems, Heather A. Eicher-Miller, Toby Jia-Jun Li, Meng Jiang, and Ronald A. Metoyer. 2024. Limitations of the LLM-as-a-judge approach for evaluating LLM outputs in expert knowledge tasks. *arXiv preprint arXiv:2410.20266*.

Quan Tu, Yanran Li, Jianwei Cui, Bin Wang, Ji-Rong Wen, and Rui Yan. 2022. MISC: A mixed strategy-aware model integrating COMET for emotional support conversation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 308–319.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, and 1 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36.

Jie Zhu, Huaixia Dou, Junhui Li, Lifan Guo, Feng Chen, Chi Zhang, and Fang Kong. 2026. Evaluating, synthesizing, and enhancing for customer support conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 35185–35194. AAAI-26 Technical Tracks 41.

A Supplementary Research Details

A.1 Literature Scope

The paper is positioned against four strands of prior work. Empathetic and support-dialogue resources motivate the strategy-aware setting (Rashkin et al., 2019; Liu et al., 2021; Tu et al., 2022; Sharma et al., 2020; Zhu et al., 2026; Budzianowski et al., 2018; Siro et al., 2024). Cultural modeling and service research motivate our interaction-level controls rather than treating culture as a single demographic attribute (Hofstede, 1984; Bond, 1996; Cao et al., 2024; Liu et al., 2024; Guo et al., 2025; Kharchenko et al., 2024; Huang and Rust, 2018). Judge-based evaluation is used with explicit caution because LLM judges are scalable but biased and context-sensitive (Gu et al., 2024; Zheng et al., 2023; Liu et al., 2023; Kumar et al., 2026; Lai et al., 2025; Szymanski et al., 2024; Samanta et al., 2026). Finally, our documentation and reliability choices follow dataset-reporting, parameter-efficient fine-tuning, and agreement-measurement literature (Bender and Friedman, 2018; Yang et al., 2024; Hu et al., 2022; James, 2026; Borin, 2022; Braylan et al., 2022; Krippendorff, 2018; Cohen, 1960; Fleiss, 1971).

A.2 Generation Card Schema

Each synthetic dialogue is generated from a structured card rather than from an unconstrained prompt. A card contains: domain, scenario, user intent, conflict level, policy constraints, user profile, cultural profile, and target strategy subset. The cultural profile is applied at the dialogue level, not as a demographic stereotype.

Control	Operational role in generation
Relationship orientation	Sets the interpersonal distance of the agent, from formal service language to warmer follow-up.
Face sensitivity	Controls whether the agent avoids direct blame and preserves the user's dignity during disagreement.
Euphemism level	Controls how negative policy information is softened, staged, or buffered.
Conflict intensity	Controls apology depth, mitigation moves, escalation handling, and solution urgency.

Table 14: Supplementary view of the cultural-control variables.

A.3 Data Example: Card-to-Dialogue Trace

Table 15 gives a compact trace of how a generated item moves from a scenario card to observable dialogue behavior. The example is translated from the Chinese synthetic data for readability. We include it to make the data construction less opaque: the cultural controls are not free-form style prompts, but intermediate variables that constrain what kinds of apologies, mitigations, and policy statements may appear.

A.4 Cultural Modeling Rationale

Our validity claim is a construct-validity argument, not a claim that a dataset can exhaustively represent Chinese culture. Let a scenario card be x, a strategy plan be s, and the cultural profile be $c = (r, f, e, k)$, where r denotes relationship orientation, f face sensitivity, e euphemism level, and k conflict intensity. The generator produces a dialogue y under a conditional distribution $p(y \mid x, s, c)$. A cultural-control variable is useful only if it satisfies three requirements.

Control relevance. Holding x and s fixed, changing c should change observable interactional behavior rather than merely changing surface length. In our design this means that high face sensitivity should increase blame avoidance and face-saving apology; higher euphemism should soften negative policy information; and higher conflict intensity should increase mitigation, expectation management, and escalation handling.

Discriminant role. Cultural controls should not collapse into the 18 strategy labels. Strategies specify *what communicative move* is intended, while cultural controls specify *how the move is realized*. For example, expectation management can be direct in a low-conflict refund inquiry, but buffered and responsibility-bounded in a high-conflict healthcare complaint. This is why the ablation separates culture removal from strategy granularity.

Empirical implication. If the construction is valid, then removing c should reduce perceived cultural fit and overall usefulness, while behavior markers should vary in the predicted directions. The paper reports this pattern: culture removal produces the largest ablation drop ($d = 1.99$), high-conflict cases show higher mitigation density, and expert calibration reaches $\kappa = 0.82$. These signals

Layer	Card field or control	Realized dialogue behavior
Scenario card	Healthcare; complaint about doctor attitude; high conflict; request for complaint channel, possible reassignment, and no delay to follow-up care.	The agent treats the case as both a service-recovery problem and a medical-continuity concern, rather than as a generic apology task.
Policy constraints	No guaranteed refund; no promise of doctor punishment; must offer bounded review or reassignment workflow.	The reply uses conditional language such as "we can submit this for review" and avoids guaranteeing a refund or disciplinary outcome.
Relationship orientation	Formal-warm service stance.	The agent addresses the user respectfully, acknowledges inconvenience, and avoids overly intimate consolation.
Face sensitivity	High.	The agent does not directly blame the doctor or the user; it frames the issue as an experience that needs verification and follow-up.
Euphemism level	Medium.	Negative information is buffered: policy limits are introduced after acknowledging the user's concern, not as a blunt refusal.
Conflict intensity	High.	The reply escalates solution specificity: record the complaint, explain the review path, offer reassignment review, and reduce paperwork anxiety.
Strategy plan	Face-saving apology; emotion validation; expectation management; follow-up closure.	The output combines apology, concrete next steps, bounded expectations, and a closing commitment to keep the user informed.

Table 15: Illustrative CCSE-CS card-to-dialogue trace. The text is translated from a synthetic Chinese healthcare-support example.

do not prove cultural completeness, but they support the narrower claim that the four controls are operationally meaningful for Chinese customer-service dialogue generation.

A.5 Example-Level Validity Checks

For each sampled card, we perform three lightweight checks before accepting the dialogue. First, the *card-faithfulness check* verifies that the answer does not contradict the domain, policy constraints, or user intent. Second, the *culture-realization check* verifies that the four controls are visible in wording choices: relational stance, face-saving behavior, indirectness of negative information, and conflict-appropriate mitigation. Third, the *strategy-coverage check* verifies that planned strategy families appear in agent turns without forcing every turn to carry a label. These checks are intentionally conservative: they reject obvious construction failures but do not replace the later ablation, expert annotation, or cross-family validation.

A.6 Prompt and Filter Details

The generation prompt has five blocks: (1) scenario metadata, (2) cultural profile, (3) target strategies, (4) user-turn plan, and (5) safety constraints. The agent is asked to keep replies bounded, actionable, and free of unverifiable guarantees. The filter stack then checks five failure classes: over-commitment, manipulative empathy, missing strategy coverage, blameful wording, and private identifiers.

Filter	Blocked behavior
PolicyGuard	Unverifiable promises, policy-violating refunds, or impossible deadlines.
EmpathySanity	Manipulative, exaggerated, or emotionally coercive language.
CoverageCheck	Dialogues that fail to express the planned strategy families.
FaceCheck	Directly blaming the user, doctor, agent, or institution in delicate cases.
PrivacyRedact	Phone numbers, names, addresses, or account-like identifiers.

Table 16: Filter stack used during construction.

A.7 Evaluation and Replication Protocol

Controlled comparisons use matched scenario cards so that only the conditioning source changes. Pairwise judging alternates AB/BA order to reduce position bias. The high-conflict replication fixes the first three user turns, regenerates only the agent side with DeepSeek-v4-Pro, and re-scores the same outputs with Claude Sonnet 4.6, GPT-4.1, and Gemini 2.5 Flash. Confidence intervals in the

construct-validity analysis use 3,000 bootstrap resamples with seed 42. The three-judge majority vote is computed per scenario card, not by averaging aggregate percentages.

Judge	Coarse	Zero	Takeaway
Claude Sonnet 4.6	**18/24 (75.0%)**	**17/24 (70.8%)**	Strong CCSE preference.
GPT-4.1	**17/24 (70.8%)**	**19/24 (79.2%)**	Strongest zero-shot separation.
Gemini 2.5 Flash	**18/24 (75.0%)**	13/24 (54.2%)	Strict on zero-shot.
Majority vote	**18/24 (75.0%)**	**17/24 (70.8%)**	Positive across families.

Table 17: Supplementary replication summary for the 24-card high-conflict test.

B Additional Diagnostic Checks

B.1 Ablation Sample Size and Judge Ceiling

The expanded ablation uses approximately 58–60 dialogues per configuration. Under a simple two-sample normal approximation, $N = 60$ per group gives roughly 80% power only for effects around $d \approx 0.51$ or larger. This is enough for the culture-removal effect ($d = 1.99$), but not for small strategy or filtering effects ($d = 0.02$ and $d = 0.26$). We therefore interpret the latter as undetected average effects, not as evidence that these modules have no operational value.

Configuration	Dim. mean	Headroom	Dims ≥4.8
Full model	**4.76**	**0.24**	**3/4 (75%)**
w/o culture	**3.46**	**1.54**	**0/4 (0%)**
Coarse-grained (6)	4.75	0.25	1/4 (25%)
With 5-fold filter	4.72	0.28	3/4 (75%)
Without filter	4.67	0.33	2/4 (50%)

Table 18: Aggregate judge ceiling diagnostic from the ablation table.

The full configuration is therefore close to the ceiling of the 1–5 judge scale. This explains why the ablation is most reliable for the large culture-removal contrast and less informative for subtle differences among already-strong configurations.

B.2 Honorific Saturation

Honorific markers are useful for verifying that the dataset uses polite service language, but they are weak for ordinal construct validity because they are present almost everywhere. Across all 9,478 dialogues, the honorific nonzero rate is 100%. Mean honorific counts are 17.93, 18.12, and 18.18 across the three relationship-orientation levels; the lexical correlation is statistically detectable but negligible ($\rho = 0.020$, $p = 0.048$), and the corresponding strategy-family correlation is null ($\rho = -0.003$, $p = 0.802$). This is why honorifics are documented here rather than used in the main construct-validity table.

C Successful Case Study

Table 19 gives a concrete success case from the high-conflict healthcare subset. The example is translated from the original Chinese outputs for readability. The fixed user turns describe a patient who felt disrespected by a doctor, asks whether a refund or doctor change is possible, and worries that paperwork will delay a follow-up visit. On this card, three-judge majority preference favors CCSE-style over the coarse baseline, and all three judges prefer it over zero-shot.

D Failure Case Analysis

We also inspect a matched card where CCSE-style control does not win. In a high-conflict telecom plan-downgrade case, the user is a long-tenured customer who wants to reduce fees but is blocked by a minimum-spend agreement. Across the three judge families, both the coarse baseline and the zero-shot baseline are preferred over the CCSE-style output. This failure is useful because it isolates a boundary of cultural control: politeness and face-saving cannot substitute for executable policy guidance.

E Known Weaknesses and Use Guidance

The strongest supported use case is training or evaluating Chinese customer-service models on culturally delicate wording, escalation handling, and strategy coverage. The dataset should not be used as evidence that synthetic agents can replace human support staff in high-stakes healthcare or financial disputes. For such settings, CCSE-CS is best treated as a pre-training or simulation resource that still requires domain-specific review, policy grounding, and deployment-time monitoring.

The most important residual risk is that cultural appropriateness is partly subjective and can sometimes over-index on warmth at the expense of operational specificity. Our expert panel, three-family

Field	Content
Scenario	Healthcare, doctor-attitude complaint, high conflict.
User summary	The patient says the doctor did not listen, asks for a feedback channel, considers refund or reassignment, and worries that complaint paperwork may delay follow-up care.
CCSE-style reply 1	Acknowledges the patient's wish to be respected, apologizes for the poor experience, records the issue, and starts internal verification without blaming the doctor.
CCSE-style reply 2	Offers bounded options: a short written explanation for refund review or reassignment to another doctor, letting the patient choose the path that fits their need.
CCSE-style reply 3	Reduces process anxiety by explaining that only a short written or online description is needed, then commits to objective feedback to the department while avoiding unrealistic guarantees.
Why it works	The response combines face-saving apology, operational next steps, blame avoidance, and trust repair. It remains concrete without promising an outcome outside the agent's authority.
Judge result	Majority preference over coarse strategy: **2/3 judges (66.7%)**. Preference over zero-shot: **3/3 judges (100%)**.

Table 19: Successful high-conflict healthcare example from the matched replication.

Field	Content
Scenario	Telecom, plan downgrade blocked by a minimum-spend agreement, high conflict.
User summary	The customer complains about three unresolved calls, frames the issue as unfair treatment of a long-tenured user, and asks whether a special exception or flexible plan is possible.
CCSE-style weakness	The response contains warm recognition of the user's status and frustration, but it relies on vague phrases such as checking for system delay, finding a way together, and seeking an answer within three working days. It also implies that old-user adjustment plans exist before verification.
Preferred baseline trait	The stronger alternatives give more concrete and bounded actions: checking the agreement end date, reviewing lower-cost eligible plans, filing an exemption/manual-review request, and setting a clearer callback or review path.
Judge result	Three-judge replication prefers the non-CCSE alternatives over CCSE-style output for both coarse and zero-shot comparisons.
Design lesson	Future filters should penalize vague special-treatment promises and require at least one domain-grounded next action whenever the conflict is policy-heavy.

Table 20: Failure case where cultural wording is less useful than concrete policy action.

judge replication, and failure-case inspection reduce this risk but do not remove it. Future versions should add larger human pairwise preference studies, domain-specific compliance audits, and real user acceptability testing under institutional review.