NICE: A Theory-Grounded Diagnostic Benchmark for Social Intelligence of LLMs

Anonymous ACL submission

Abstract

As large language models (LLMs) are increasingly applied in social contexts such as emotional companionship and customer service, measuring their social intelligence has become critical to the quality and safety of human–AI interaction. However, existing social intelligence benchmarks lack a unified framework that organizes social abilities into a unified structure, and therefore cannot enable fine-grained diagnosis. To build the first holistic diagnostic evaluation grounded in social theory, we first construct a social intelligence framework through a literature review and multi-stage expert validation guided by psychometric principles. The resulting framework includes 4 categories and 11 dimensions, each further specified by fine-grained capability facets. Building on this framework, we introduce **NICE** (**N**orm, **I**nteraction, **C**ognition, **E**xperience), a diagnostic benchmark of 137 items operationalized through representative Chinese contexts. Across 5 frontier LLMs and a human reference group, models score higher in aggregate accuracy yet show a consistent weakness in Communication, which the framework localizes to 3 specific capability facets: multi-turn communication, nonverbal communication, and synchrony. NICE thus reframes social intelligence evaluation toward theory-grounded diagnosis of socially consequential weaknesses in LLMs.

1 Introduction

Social intelligence, the capacity to understand, integrate into, and adapt to social environments, is a core aspect of intelligent social behavior (Thorndike, 1920; Gardner, 2011; Salovey and Mayer, 1990). As large language models (LLMs) are increasingly deployed in socially intensive contexts, evaluating their social intelligence has become a central research topic (Dmonte et al., 2024; Hou et al., 2025; Wang et al., 2024b). Existing benchmarks fall broadly into two lines. One targets a single, well-defined social ability, such as theory of mind, emotion understanding, or moral judgment (Wu et al., 2023; Zadeh et al., 2019; Sabour et al., 2024; Iyer et al., 2026; Yu et al., 2024), grounding each task in a specific construct for clear interpretability within that ability. The other places models in open-ended, multi-turn scenarios (Wang et al., 2024b), trading control for higher ecological validity.

Despite this progress, few benchmarks evaluate social intelligence as a whole. Single-ability benchmarks measure isolated constructs, with no shared framework to show how strengths and weaknesses are distributed across the broader social-intelligence space. Simply aggregating them is also insufficient, because their constructs, formats, metrics, and validation standards are difficult to align. Open-ended benchmarks cover many abilities at once, but entangle them within a single scenario, so failures cannot be traced to specific abilities. Either way, models are reduced to overall scores that show how much they get right but not where they fall short, offering little insight into the structure of social abilities needed for diagnosis and improvement.

To address these gaps, we propose **NICE** (**N**orm, **I**nteraction, **C**ognition, **E**xperience), a theory-grounded diagnostic benchmark for artificial social intelligence. To our knowledge, NICE is the first to apply measurement principles throughout framework construction, item development, and validation. We first construct a social intelligence framework through a literature review and multi-stage expert validation. The framework organizes social intelligence into 4 categories and 11 dimensions, each refined into fine-grained capability facets, with structural weights over the categories and dimensions estimated by the Analytic Hierarchy Process (AHP; Saaty, 1987). Guided by this framework, we then build a human-written item pool through repeated review and revision, aligning each

of the final 137 items with a single capability facet so that item performance provides interpretable evidence for the targeted facet. Each item is posed as a closed-form ranking task, where the model orders candidate responses from socially optimal to boundary-violating, testing not only whether it selects the best response but also whether it respects behavioral and normative boundaries. Our contributions are as follows:

- To our knowledge, NICE is the first systematic and comprehensive diagnostic benchmark for artificial social intelligence. Built on a theory-grounded framework, NICE organizes social intelligence into 4 categories, 11 dimensions and 34 facets, enabling evaluation beyond single abilities or domain-specific tasks.
- NICE enables facet-level diagnosis. Each item is tied to one predefined capability facet, so model errors can be traced to specific weaknesses rather than being absorbed into dimension-level or overall scores.
- We developed NICE through a full psychometric pipeline rather than as a simple task collection. By applying measurement principles across framework construction, item development, and validation, NICE strengthens framework validity and makes its diagnostic results more interpretable.

2 Related Work

Existing work on benchmarking the social capabilities of LLMs can be broadly divided into two categories: one focuses on specific dimensions of social capability, the other emphasizes situated or interactive evaluation. While these works often apply psychometric or quality-control procedures at the dimension or task level, such partial checks do not ensure that benchmark scores map to specific and interpretable constructs. NICE therefore applies measurement principles throughout the full pipeline, from holistic framework construction to item development and validation. Table 1 compares NICE with representative social benchmarks in construction, validation, and diagnostic design.

For benchmarks targeting specific dimensions of social capabilities, representative examples include Social IQa (Sap et al., 2019), which centers on social commonsense reasoning, as well as HI-ToM (Wu et al., 2023) and ToMBench (Chen et al., 2024b), which extend benchmark design to higher-order mental-state inference. Other studies focus on emotion understanding and empathy, such as EmoBench (Sabour et al., 2024), or moral judgment and value alignment, such as CMoralEval (Yu et al., 2024). Most of these benchmarks treat social intelligence as a collection of isolated tasks. They lack a unified capability structure that can integrate different dimensions and reflect their internal dependencies. In contrast to these task-driven benchmarks, NICE is built upon a holistic theoretical framework of social intelligence whose component weights are empirically calibrated.

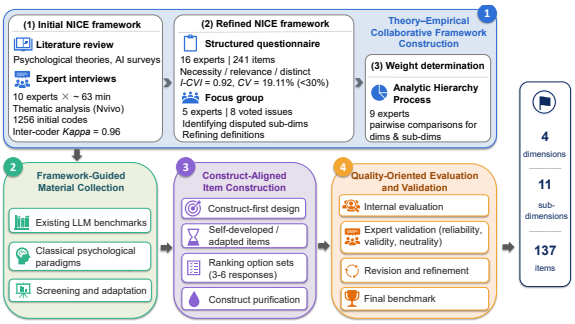

Figure 1: NICE construction pipeline. The benchmark is built through framework construction, material collection, item construction, item evaluation and validation.

Another line of work aims to assess model capabilities in interactive settings with higher ecological validity. For example, SocialBench (Chen et al., 2024a) examines role-playing performance in individual and group interactions; RoleBench (Wang et al., 2024a) benchmarks role-playing ability at a more fine-grained level; and SOTOPIA (Zhou et al., 2024) evaluates model behavior in dynamic social scenarios through multi-turn interaction environments. However, the data construction of these benchmarks often relies on automated generation pipelines or uncurated corpus crawling. Without rigorous psychometric constraints, such construction paradigms can easily introduce contextual noise and construct overlap across dimensions, thereby limiting their diagnostic precision for deeper model deficiencies. By contrast, NICE performs construct purification through multiple rounds of expert review and iteration, maximizing the alignment between benchmark results and the underlying capability constructs.

3 NICE

We develop NICE as a social intelligence benchmark with clearly defined constructs and fine-

Benchmark	Target	Framework-guided	Data generation	Data validation (per item)	Format	Granularity
Social IQa	Social commonsense	No	ATOMIC-seeded crowdsourcing	8 workers (2 rounds) + artifact control	MCQ	Task
ToMBench	Theory of Mind	Yes	Human-written	3 workers (2–3 rounds)	MCQ	Dimension
EmoBench	Emotional intelligence	No	LLM-inspired, human-written	4 workers	MCQ	Dimension
CMoralEval	Moral judgment	No	Corpus-derived, LLM-assisted	2 annotators + 1 expert	MCQ	Dimension
RoleBench	Role-playing	No	Script-derived, LLM-generated	Sampled review; unclear number	Instruction-response	Character
SocialBench	Social interaction	No	Multi-source, LLM-assisted	3 annotators	MCQ + open generation	Dimension
SOTOPIA	Social interaction	No	LLM-generated, dataset-inspired	Author review (unclear number) + LLM evaluation	Multi-turn interaction	Episode
HEART	Emotional support	Yes	ESConv-derived, adversarial variants	5 raters	Pairwise comparison	Dimension
NICE	**Holistic ASI**	**Yes**	**Human-written**	**12 evaluators (3 rounds)**	**Ranking**	**Facet**

Table 1: Comparison of representative social intelligence benchmarks and NICE, highlighting differences in full-pipeline measurement principles and diagnostic granularity. The mapping of representative benchmarks to NICE categories is shown in the left panel of Figure 2. MCQ = multiple-choice question; ASI = artificial social intelligence. Granularity denotes the finest level at which each benchmark is designed to support diagnostic interpretation.

grained diagnostic power. Its construction follows four stages: framework construction, material collection, item construction, and item evaluation and validation (Figure 1). This design substantially reduces construct contamination and ensures that each benchmark item points more clearly to a specific social capability dimension and its corresponding facets.

3.1 Framework Construction

Following the methodological logic of psychometric scale development and expert-consensus research, we adopt a three-stage procedure to construct the framework, combining a top-down systematic literature review with a bottom-up Delphi method (Process 1 in Figure 1): Candidate constructs are **first** defined from theoretical and empirical evidence, **then** refined through structured expert judgment, and **finally** operationalized for the target evaluation context (Chu and Hwang, 2008; Murry Jr and Hammons, 1995). Across the framework-construction process, 23 experts contributed to one or more stages, yielding 40 stage-level participations across interviews, structured ratings, focus-group discussion, and AHP weighting. All the experts in this process were recruited according to predefined eligibility criteria related to disciplinary background, methodological familiarity, and experience in psychology, social intelligence, or human–AI interaction (Appendix B).

Initial Framework Construction We center our framework on social information processing theory (Crick and Dodge, 1994), and integrate 8 classical theories of human social intelligence (Thorndike, 1920; Guilford, 1967; Gardner, 2011; Baron-Cohen, 1989; Salovey and Mayer, 1990; Crick and Dodge, 1994; Goleman, 1995; Rose-Krasnor, 1997), as well as reviews on AI social intelligence from the past 5 years (Fan et al., 2022; Langley et al., 2022; Niu et al., 2025; Williams et al., 2022). In parallel, we conducted semi-structured in-depth interviews with 10 experts to elicit their independent understanding of AI social intelligence and to refine the literature-derived initial framework. We then applied thematic analysis (Braun and Clarke, 2006) to refine 1,256 initial codes. The inter-coder agreement *Kappa* = 0.96, indicating good reliability. Based on the literature review and expert interviews, we develop an initial social intelligence framework that incorporates both classical dimensions of human social intelligence and characteristics specific to LLMs.

Structured Ratings and Focus-group Discussion To refine the initial framework derived from the literature review and expert interviews, we invited 16 experts to evaluate each component in the initial framework on a 4-point Likert scale in terms

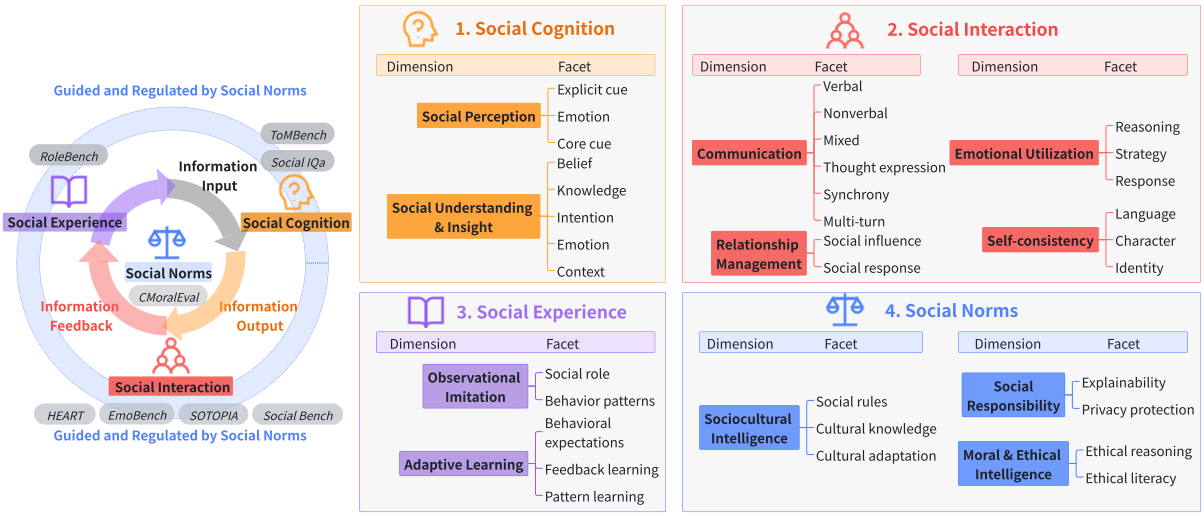

Figure 2: The framework of **NICE** (**N**orm, **I**nteraction, **C**ognition, and **E**xperience). The left panel shows the logical relations among the 4 categories, with *italicized text* indicating representative benchmarks in Table 1 within NICE category; the right panel presents the 11 dimensions and 34 facets.

of necessity, relevance, and discriminability. The results showed that the average coefficient of variation (CV) of expert ratings is 19.11%, substantially below the 30% threshold. The item-level content validity index (I-CVI) reached 0.92, indicating good structural stability and expert acceptance of the model (Polit and Beck, 2006). For the dimensions that still show disagreement, we further organize a focus group with 5 experts and conducted anonymous voting on 8 issues, merging highly overlapping dimensions and refining their operational definitions.

AHP Weighting A total of 9 experts conducted pairwise comparisons of the importance of each component; the final structural weights (Appendix A) were derived from the 7 experts whose dimension-level Consistency Ratio was below 0.1. This weighting scheme not only reflects the relative importance of different capability dimensions within overall social intelligence but also enables NICE to identify capability deficits with greater diagnostic significance and higher priority for improvement, thereby informing subsequent model training and alignment.

Final Framework The final framework for social intelligence consists of 4 categories—social **Cognition**, social **Interaction**, social **Experience**, social **Norm**—and 11 fine-grained dimensions (Figure 2). Aligned with the logic of social information processing (Crick and Dodge, 1994), we conceptualize social intelligence as a process involving information input, information output, and information feedback. Information input requires perceiving, interpreting, and inferring socially relevant cues, corresponding to **social cognition**; information output requires selecting appropriate communicative, emotional, relational, and behavioral strategies, corresponding to **social interaction**; and information feedback requires learning from observed behavior, interaction outcomes, and feedback signals, corresponding to **social experience**. **Social norm** further provides the background knowledge and regulatory constraints that guide the entire process, including sociocultural rules, moral constraints, and responsibilities. Together, these dimensions specify how LLMs understand social information, act in interaction, learn from feedback, and regulate behavior according to normative knowledge. Each dimension is associated with clear facets and weight, where the facets define the boundaries of the target capability and guide item design, and the weight reflects its relative importance.

3.2 Material Collection and Item Construction for NICE

Material Collection NICE does not directly extract items at scale from real-world corpora or ready-made scenarios. Real social situations often involve multiple cues and irrelevant factors, and direct extraction may cause a single sample to cover multiple capability facets, thereby undermining result interpretation. In such cases, when

a model fails, it becomes difficult to determine what specific error caused the failure, which conflicts with our goal of fine-grained model diagnosis. Therefore, we refer to materials from existing LLM benchmarks and classical psychological paradigms primarily as development references (Process 2 in Figure 1). The benchmarks are used to identify adaptable task formats for social intelligence, while the psychological paradigms offer well-established measurement paradigms for social cognition and social interaction. In total, we identified 18 adaptable benchmarks from 42 candidate papers and compiled 43 relevant psychological tests or experimental paradigms as development references.

Item Construction Building on the theoretical framework and reference materials, we developed 137 items aligned with the framework above (Process 3 in Figure 1). Two researchers with 7–8 years of psychological research experience identified each target dimension and capability facet (Appendix A), and then designed the scenario and decision question accordingly (Appendix B.3 describes item-developer eligibility). When existing materials had ambiguous construct boundaries or insufficient coverage, they were adapted or replaced with newly created items to ensure facet-specific alignment. To improve diagnostic granularity, NICE adopts a closed-form ranking task (Figure 3), in which each item contains 3–6 candidate responses forming a controllable quality gradient: an optimal response that fits the situation, interaction goal, and social norms; a suboptimal response that is partially reasonable but limited in understanding, expression, or normative appropriateness; and a worst response that ignores key information or violates the interaction goal, relational boundary, or normative requirement. This design supports boundary-sensitive diagnosis at the facet level.

3.3 Item Evaluation and Validation for NICE

After the initial item pool was constructed, NICE underwent a three-round item evaluation and revision process (Process 4 in Figure 1). In the first round, 2 trained interdisciplinary evaluators conducted a small-scale internal evaluation to detect construct drift, ambiguous wording, unstable response gradients, and potential neutrality issues. In the second round, the revised item pool was independently evaluated by 10 additional trained interdisciplinary evaluators with expertise in psychology and human–AI interaction. None of the

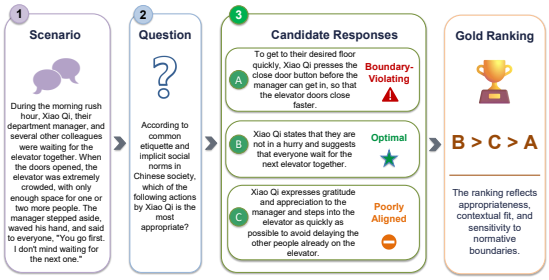

Figure 3: Example of a ranking item. Given a scenario, participants rank 3 candidate responses from optimal to worst. The gold ranking reflects appropriateness, contextual fit, and sensitivity to normative boundaries.

12 evaluators was involved in item development, and all were blinded to subsequent model evaluation results (evaluator recruitment and screening criteria are presented in Appendix B.3). In the third round, items revised after external validation were returned to the evaluator(s) who assigned low scores for confirmation and re-scoring.

Evaluators assessed each item along reliability, validity, and neutrality. Reliability concerned wording clarity, response-gradient and reference-answer accuracy; validity concerned construct alignment; and neutrality concerned bias and unnecessary value assumptions. Each criterion was rated on a 5-point scale, with 3.5 as the predefined retention threshold. Items scoring at least 3.5 on all 3 criteria were retained, whereas items receiving any sub-threshold score were revised according to the corresponding evaluator feedback and re-evaluated until they met the threshold. After this iterative process, all retained items met the predefined quality threshold and were included in the final benchmark (Appendix C1). This validation checks every item through multiple evaluators and iterative revision, emphasizing construct quality beyond benchmarks that focus on scale, format, or interactive settings (Section 2).

3.4 Final Benchmark

The final NICE is a theory-grounded diagnostic benchmark for social intelligence. It covers 4 categories, 11 dimensions, and 137 benchmark items (Appendix A). Each item is mapped to a clearly defined capability facet and paired with a structured set of candidate responses and an expert-defined gold ranking. Overall, NICE emphasizes item-to-construct alignment and systematic validation aligned with a framework. It is designed to pro-

vide comparable, interpretable, and fine-grained diagnostic signals for identifying weaknesses in the social intelligence of models.

4 Evaluation Protocol

Evaluated Models We evaluate 5 frontier LLMs: GPT-5.5 (OpenAI, 2026), Claude-Opus-4.7 (Anthropic, 2026), Gemini-3.1-pro-preview (Google DeepMind, 2026), DeepSeek-v4-pro (DeepSeek-AI, 2026), and Qwen3.6-plus (Qwen Team, 2026). Each model is accessed via its official first-party API under zero-shot, greedy decoding settings (temperature $= 0$, top_p $= 1$). To assess residual non-determinism, each model runs the full 137-item benchmark 3 times independently (2,055 total calls). Exact model snapshots, API call dates, and prompt templates are detailed in Appendix D.

Task Format NICE adopts a closed-form ranking task (Figure 3). For each item, models receive an input consisting of a scenario, a question, and candidate responses. Models are required to rank the 3 options from best to worst. During testing, 1 candidate response is sampled from each of the optimal, suboptimal, and worst response pools using a fixed random seed. This design is consistent with the hierarchical option-set construction used in NICE and more directly evaluates models' sensitivity to social intelligence boundaries.

Metrics A prediction is correct only if the full permutation over the three candidate responses matches the gold ranking; partial agreement receives no credit. Model outputs are parsed by a deterministic regex extracting the first three distinct digits in 1, 2, 3. We report (i) overall accuracy, (ii) per-dimension accuracy, and (iii) cross-run sample standard deviation as a stability indicator.

Human Baseline To establish human reference performance on NICE, we recruited 14 adult native Chinese speakers with undergraduate-level or higher education (referring to Hou et al., 2025; Iyer et al., 2026). Participants were asked to complete the same ranking items as the evaluated LLMs, and their average accuracy was used as the human baseline. For in-depth analysis, we further presented the mean accuracy of the top 3 participants within each dimension. No additional tutorials, demonstrations, or task-specific feedback were provided during the experiment, ensuring that humans and models were evaluated under comparable conditions.

5 Results and Analysis

5.1 Main Results

Overall Performance Table 2 shows the main results. LLMs show higher alignment with expert-defined gold rankings than the human baseline on average (LLMs: $M = 0.751$ $[0.734, 0.769]$; humans: $M = 0.704$ $[0.687, 0.719]$; Welch's $t(26.9) = 3.80$, $p < .001$, $d = 1.41$), a gap of approximately 4.7%. Gemini-3.1-pro-preview ($M = 0.788$) and GPT-5.5 ($M = 0.786$) ranked highest and were within 2.0% of each other, while Claude-Opus-4.7 ranked lowest ($M = 0.710$). Across dimensions, models perform best on D4 and worst on D3. The weakness aligns with humans, who are also lowest on D3; the strength does not, as humans peak on D5. Humans descriptively outperform models on D3 and D9, with the D3 gap reaching 9.0% (CI $[-0.171, -0.008]$). Models outperform humans on D10, D6, and D4 by 17.4, 13.9, and 12.8% respectively, with all three gaps supported by bootstrap CIs excluding zero (D10: $[+0.098, +0.254]$; D6: $[+0.071, +0.203]$; D4: $[+0.076, +0.186]$).

Dimension-level Model–human Gaps The LLM advantage is concentrated in specific dimensions rather than evenly distributed (Table 3). Of the eleven dimensions, six show bootstrap CIs that exclude zero: LLMs score robustly higher on Social Responsibility, Self-consistency, Emotional Utilization, Social Perception, and Adaptive Learning (gaps of +10.5 to +17.3 pp; full CIs in Table 3), while the remaining five dimensions show no detectable difference. Going in the other direction, D3 Communication is the only dimension where humans robustly outperform LLMs—a gap of 9.0 pp whose CI just excludes zero ($[-0.170, -0.008]$), motivating the targeted analysis in Section 5.2.

5.2 D3 (Communication) as a Consistent Diagnostic Weakness of Frontier LLMs

Among the 11 dimensions, Communication (D3) emerges as the lowest-scoring dimension for all 5 LLMs we evaluated. To illustrate the diagnostic value of NICE beyond aggregate accuracy, we select D3 as a representative case for in-depth analysis. We examine whether its low performance reflects a structured weakness rather than an isolated artifact, using the following 5 complementary perspectives.

Model	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	Weighted All	All [95% CI]
GPT-5.5	0.750	0.889	0.361	1.000	0.821	0.806	0.556	0.889	0.750	1.000	0.813	0.760	0.786 $[0.720, 0.849]$
Qwen3.6-plus	0.778	0.750	0.444	0.917	0.821	0.667	0.611	0.806	0.667	0.806	0.708	0.714	0.725 $[0.655, 0.791]$
Gemini-3.1-pro-preview	0.806	0.917	0.389	1.000	0.923	0.667	0.722	0.833	0.750	0.972	0.708	0.781	0.788 $[0.723, 0.852]$
DeepSeek-V4-pro	0.806	0.806	0.444	0.917	0.846	0.806	0.611	0.750	0.611	0.806	0.792	0.734	0.747 $[0.686, 0.808]$
Claude-Opus-4.7	0.806	0.667	0.500	0.944	0.846	0.667	0.694	0.667	0.583	0.917	0.563	0.705	0.711 $[0.635, 0.783]$
Human	0.667	0.815	0.518	0.827	0.846	0.583	0.607	0.685	0.714	0.726	0.732	0.701	0.704 $[0.660, 0.746]$

Table 2: Accuracy of frontier LLMs and human participants on the NICE. The *Weighted All* column indicates the accuracy after weighting across 11 dimensions. 95% bootstrap CIs in the *All* column are computed by resampling items ($n = 137$, 10,000 resamples). D1 = Social Perception; D2 = Social Understanding & Insight; D3 = Communication; D4 = Emotional Utilization; D5 = Relationship Management; D6 = Self-consistency; D7 = Observational Imitation; D8 = Adaptive Learning; D9 = Sociocultural Intelligence; D10 = Social Responsibility; D11 = Moral & Ethical Intelligence.

		Human ($n = 14$)			**LLM** ($n = 5$)			**Diff. (LLM−H)**
ID	**Dimension**	M	SD	95% CI	M	SD	95% CI	95% CI
D1	Social Perception	0.667	0.122	$[0.601, 0.726]$	0.789	0.062	$[0.761, 0.822]$	$[+0.053, +0.192]$
D2	Social Understanding	0.815	0.088	$[0.774, 0.863]$	0.806	0.112	$[0.750, 0.861]$	$[-0.081, +0.059]$
D3	Communication	0.518	0.139	$[0.446, 0.583]$	0.428	0.083	$[0.389, 0.467]$	$[-0.170, -0.008]$
D4	Emotional Utilization	0.827	0.095	$[0.780, 0.869]$	0.956	0.053	$[0.928, 0.978]$	$[+0.075, +0.186]$
D5	Relationship Management	0.846	0.080	$[0.802, 0.885]$	0.851	0.054	$[0.826, 0.877]$	$[-0.043, +0.053]$
D6	Self-Consistency	0.583	0.098	$[0.536, 0.631]$	0.722	0.087	$[0.678, 0.767]$	$[+0.073, +0.203]$
D7	Observational Imitation	0.607	0.106	$[0.554, 0.661]$	0.639	0.081	$[0.600, 0.678]$	$[-0.035, +0.097]$
D8	Adaptive Learning	0.685	0.131	$[0.619, 0.750]$	0.789	0.088	$[0.744, 0.833]$	$[+0.023, +0.181]$
D9	Sociocultural Intelligence	0.714	0.149	$[0.637, 0.792]$	0.672	0.086	$[0.633, 0.717]$	$[-0.129, +0.042]$
D10	Social Responsibility	0.726	0.124	$[0.661, 0.786]$	0.900	0.096	$[0.850, 0.944]$	$[+0.098, +0.251]$
D11	Moral & Ethical Intelligence	0.732	0.111	$[0.670, 0.781]$	0.717	0.097	$[0.667, 0.762]$	$[-0.085, +0.062]$
All	Average	0.704	0.032	$[0.687, 0.719]$	0.751	0.036	$[0.734, 0.769]$	$[+0.014, +0.081]$

Table 3: Per-dimension accuracy of human participants and LLMs on NICE, with means (M), standard deviations (SD), and 95% percentile bootstrap confidence intervals (10,000 resamples). Human $n = 14$ participants; LLM $n = 15$ models. The rightmost column gives the bootstrap CI of the LLM−human difference, and dimensions whose CIs exclude zero are highlighted in grey. D3 is the only dimension where humans outperform LLMs; 5 dimensions show a robust LLM advantage.

Universal Weakness Across LLMs Figure 4A shows the per-dimension accuracy of the 5 evaluated LLMs, compared with both the average human reference and human Top-3. D3 is the lowest-scoring dimension for both LLMs and human, with a 5-model mean of 0.428 (95% CI $[0.389, 0.467]$), falling 23.9% below the human Top-3. Crucially, for each of the 5 models the 95% bootstrap CI of D3 accuracy lies entirely below the CI of that model's non-D3 accuracy. This pattern holds across all 5 models, arguing against any single-model explanation.

Decoupling from Overall Accuracy Across the 5 models, D3 rankings diverge from their overall benchmark rankings. The 2 strongest overall performers (Gemini-3.1-pro-preview and GPT-5.5) score lowest on D3, while the model that scores highest on D3 ranks last overall. This decoupling indicates that the ability to navigate communicative exchange is not captured by the aggregate signal that drives overall benchmark accuracy. Communication competence thus appears to constitute a relatively independent skill, and current scaling or alignment practices may not reliably improve it.

Localization to Specific Communication Facets D3 comprises 6 capability facets, and LLM deficits are unevenly distributed across them (Figure 4B). The largest gaps occur in *Multi-turn communication*, where LLMs lag human by 52.4% (CI $[-75.7, -28.6]$). Gaps in *Nonverbal communication* (24.8%) and *Synchrony* (10.5%) are directionally consistent but treated as exploratory given the small number of items per facet. These weaknesses are also highly model-dependent. For example, on *Multi-turn communication*, Claude-Opus-4.7 answers every item correctly, whereas Qwen3.6-Plus answers none correctly ($SD = 0.35$). On *Nonverbal communication*, Qwen3.6-Plus and Gemini-3.1-pro-preview answer every item correctly, while Claude-Opus-4.7 and GPT-5.5 fail every item ($SD = 0.45$). This variance suggests that the underlying capabilities are un-

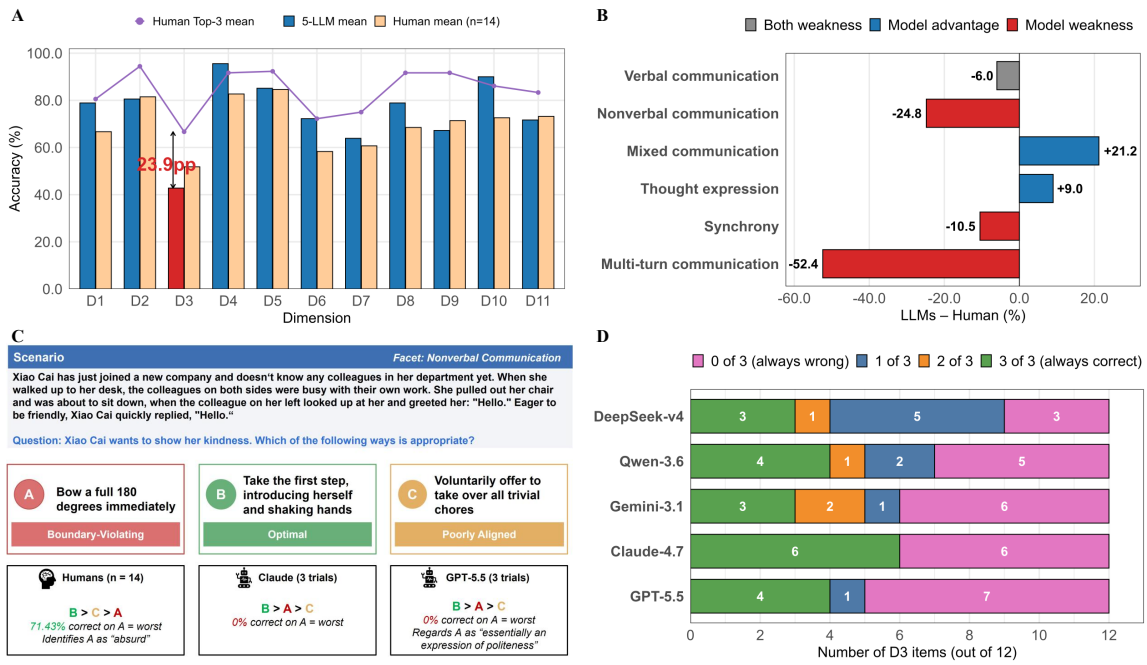

Figure 4: Evidence for Communication as a consistent weakness of frontier LLMs. (A) Dimension-level accuracy for the 5-LLM mean, human mean, and Human Top-3; D3 shows the lowest model accuracy and a 23.9-point gap from Human Top-3. (B) D3 facet-level LLM–human gaps, concentrated in *multi-turn communication*, *nonverbal communication*, and *synchrony*. (C) Case study showing that LLMs under-penalize excessive, boundary-violating deference. (D) Across-run consistency, where D3 items are consistently failed.

evenly represented across model families rather than uniformly absent. By contrast, LLMs exceed humans on *Mixed communication* (LLMs - humans = 21.2%, CI $[+0.5, +39.0]$) and *Thought expression* (9.0%), which likely rely more on declarative knowledge of communicative conventions.

Case Study: Excessive Politeness Figure 4C illustrates a *Nonverbal communication* failure in a first-meeting scenario. All models and humans correctly identify Option B as the optimal action, but they diverge on the boundary-violating option: 71.43% of humans ranked Option "bowing at 180 degrees" as the worst choice, whereas Claude-Opus-4.7 and GPT-5.5 never did so across 3 independent trials, consistently placing it second. Follow-up explanations show that a human participant viewed the exaggerated bow as a implausible violation of real-world conventions, while GPT-5.5 interpreted it as an expression of politeness. This case shows that LLMs may overvalue explicit deference and under-penalize contextually inappropriate behavior, revealing boundary failures that would be hidden if only the optimal choice were evaluated.

Item-level Stability Across Repeated Runs Figure 4 D shows the distribution of D3 items by the number of correct responses across each model's three independent runs. Among the 60 (model, item) cells, 78.3% fall at the two extremes, always correct or always incorrect, compared with 25% expected under an i.i.d. Binomial null ($\chi^2(3) = 94.49$, $p < .001$). This bimodality indicates that D3 failures reflect systematic patterns in model behavior rather than random sampling noise — when a model fails a D3 item, it fails it reliably.

6 Conclusion

We introduce NICE, a systematic benchmark for diagnosing LLM social intelligence. Built through a psychometric pipeline, NICE organizes social intelligence into 4 categories and 11 dimensions and ties each item to one predefined capability facet. This design turns evaluation into diagnosis by tracing model errors to specific social-intelligence weaknesses. Experiments on frontier LLMs and a human reference group show that, despite higher aggregate model accuracy, LLMs exhibit consistent weaknesses in Communication, especially in facets requiring sensitivity to subtle interactional cues and boundaries. NICE provides a theory-grounded diagnostic tool for understanding and improving LLMs in real-world social interaction.

7 Limitations

Static Text-based Benchmark for Social Intelligence NICE currently evaluates the social intelligence of LLMs through static, text-based scenarios. We design the scenarios to approximate real-life social situations, and use construct alignment and expert validation to ensure that each item targets a specific capability facet. However, NICE remains a static, text-based benchmark and therefore cannot fully capture the complexity and dynamics of real social interaction. To support precise diagnosis, NICE deliberately controls scenario descriptions so that each item can be mapped to a clear capability facet. In real-world social settings, social cognition, interactional response, emotion regulation, norm compliance, and feedback adaptation often unfold together across continuous interaction. As a result, the current version of NICE is better suited to diagnosing localized weaknesses in specific social capability dimensions than to evaluating how models coordinate multiple social abilities in complex dynamic contexts. Future work can extend NICE to more complex social scenarios and interactive settings, enabling a more comprehensive assessment of LLM social intelligence in real-world environments.

Cultural Scope and Generalizability NICE items are developed primarily within a Chinese cultural context. Although the framework draws on cross-cultural theories of social intelligence, the operationalization of certain social norms may not generalize across cultures. Therefore, the results should be interpreted with this scope in mind. Future work should extend NICE to broader contexts and larger samples, and examine the cross-cultural robustness of its item design and diagnostic inferences.

8 Ethical Considerations

We accessed all evaluated LLMs through their official APIs in compliance with each provider's terms of service, and used the outputs solely for academic evaluation. We further caution that NICE is intended as a diagnostic tool for identifying capability weaknesses, and aggregate scores should not be taken as endorsements of model deployment in socially sensitive applications such as companionship for vulnerable populations, since benchmark performance does not guarantee safe interaction in dynamic real-world settings.

To establish the human baseline for the NICE benchmark, we recruited participants who are native Chinese speakers. This sampling decision was based on practical constraints and should not be interpreted as reflecting any bias or unfairness toward individuals of any race, nationality, or other personal characteristics. All participants provided informed consent and participated voluntarily. The experiment did not require participants to provide sensitive or personally identifiable information. Participants received either course credit or monetary compensation promptly after completing the experiment.

References

Anthropic. 2026. Introducing Claude Opus 4.7. Technical report, Anthropic. Accessed: 2026-04-29.

Simon Baron-Cohen. 1989. The autistic child's theory of mind: A case of specific developmental delay. *Journal of child Psychology and Psychiatry*, 30(2):285–297.

Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101.

Hongzhan Chen, Hehong Chen, Ming Yan, Wenshen Xu, Gao Xing, Weizhou Shen, Xiaojun Quan, Chenliang Li, Ji Zhang, and Fei Huang. 2024a. Socialbench: Sociality evaluation of role-playing conversational agents. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2108–2126.

Zhuang Chen, Jincenzi Wu, Jinfeng Zhou, Bosi Wen, Guanqun Bi, Gongyao Jiang, Yaru Cao, Mengting Hu, Yunghwei Lai, Zexuan Xiong, and 1 others. 2024b. Tombench: Benchmarking theory of mind in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15959–15983.

Hui-Chun Chu and Gwo-Jen Hwang. 2008. A delphi-based approach to developing expert systems with the cooperation of multiple experts. *Expert systems with applications*, 34(4):2826–2840.

Nicki R Crick and Kenneth A Dodge. 1994. A review and reformulation of social information-processing mechanisms in children's social adjustment. *Psychological bulletin*, 115(1):74.

DeepSeek-AI. 2026. DeepSeek-V4 Technical Report. *arXiv preprint*. Accessed: 2026-04-29.

Alphaeus Dmonte, Tejas Arya, Tharindu Ranasinghe, and Marcos Zampieri. 2024. Towards generalized offensive language identification. In *International Conference on Advances in Social Networks Analysis and Mining*, pages 271–286. Springer.

Lifeng Fan, Manjie Xu, Zhihao Cao, Yixin Zhu, and Song-Chun Zhu. 2022. Artificial social intelligence: A comparative and holistic view. *CAAI Artificial Intelligence Research*, 1(2):144–160.

Howard Gardner. 2011. *Frames of mind: The theory of multiple intelligences*. Basic books.

Daniel Goleman. 1995. Emotional intelligence bantam books. *New York*.

Google DeepMind. 2026. Gemini 3.1: Technical Report. Technical report, Google. Accessed: 2026-04-29.

Joy Paul Guilford. 1967. The nature of human intelligence.

Guiyang Hou, Wenqi Zhang, Yongliang Shen, Zeqi Tan, Sihao Shen, and Weiming Lu. 2025. Egosocialarena: Benchmarking the social intelligence of large language models from a first-person perspective. *Preprint*, arXiv:2410.06195.

Laya Iyer, Kriti Aggarwal, Sanmi Koyejo, Gail Heyman, Desmond C Ong, and Subhabrata Mukherjee. 2026. Heart: A unified benchmark for assessing humans and llms in emotional support dialogue. *arXiv preprint arXiv:2601.19922*.

Christelle Langley, Bogdan Ionut Cirstea, Fabio Cuzzolin, and Barbara J Sahakian. 2022. Theory of mind and preference learning at the interface of cognitive science, neuroscience, and ai: A review. *Frontiers in artificial intelligence*, 5:778852.

John W Murry Jr and James O Hammons. 1995. Delphi: A versatile methodology for conducting qualitative research. *The review of higher education*, 18(4):423–436.

Ruowen Niu, Jiaxiong Hu, Siyu Peng, Caleb Chen Cao, Chengzhong Liu, Sirui Han, and Yike Guo. 2025. Scenario, role, and persona: A scoping review of design strategies for socially intelligent ai agents. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–9.

OpenAI. 2026. GPT-5.5 System Card. Technical report, OpenAI. Accessed: 2026-04-29.

Denise F Polit and Cheryl Tatano Beck. 2006. The content validity index: are you sure you know what's being reported? critique and recommendations. *Research in nursing & health*, 29(5):489–497.

Qwen Team. 2026. Qwen3.6-Plus: Advancing Large Language Models via Continuous Pre-training and Alignment. Alibaba Group Official Blog. Accessed: 2026-04-29.

Linda Rose-Krasnor. 1997. The nature of social competence: A theoretical review. *Social development*, 6(1):111–135.

Roseanna W Saaty. 1987. The analytic hierarchy process—what it is and how it is used. *Mathematical modelling*, 9(3-5):161–176.

Sahand Sabour, Siyang Liu, Zheyuan Zhang, June Liu, Jinfeng Zhou, Alvionna Sunaryo, Tatia Lee, Rada Mihalcea, and Minlie Huang. 2024. Emobench: Evaluating the emotional intelligence of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5986–6004.

Peter Salovey and John D Mayer. 1990. Emotional intelligence. *Imagination, cognition and personality*, 9(3):185–211.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social iqa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 4463–4473.

Edward L Thorndike. 1920. Intelligence and its uses. *Harper's magazine*, 140:227–235.

Noah Wang, Zy Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Jian Yang, and 1 others. 2024a. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 14743–14777.

Ruiyi Wang, Haofei Yu, Wenxin Zhang, Zhengyang Qi, Maarten Sap, Yonatan Bisk, Graham Neubig, and Hao Zhu. 2024b. Sotopia-π: Interactive learning of socially intelligent language agents. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12912–12940.

Jessica Williams, Stephen M Fiore, and Florian Jentsch. 2022. Supporting artificial social intelligence with theory of mind. *Frontiers in artificial intelligence*, 5:750763.

Yufan Wu, Yinghui He, Yilin Jia, Rada Mihalcea, Yulong Chen, and Naihao Deng. 2023. Hi-tom: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10691–10706.

Linhao Yu, Yongqi Leng, Yufei Huang, Shang Wu, Haixin Liu, Xinmeng Ji, Jiahui Zhao, Jinwang Song, Tingting Cui, Xiaoqing Cheng, and 1 others. 2024. Cmoraleval: A moral evaluation benchmark for chinese large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11817–11837.

Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. 2019. Social-iq: A question answering benchmark for artificial social

intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8807–8817.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, and 1 others. 2024. Sotopia: Interactive evaluation for social intelligence in language agents. In *International Conference on Learning Representations*, volume 2024, pages 40975–41019.

A Information of the NICE Benchmark

Category	Sub-Dimension (Number of Items)	Facet	Dimension Weight	Number of Social Scenarios	Distribution of Item Development
1 Social Cognition	1-1 Social Perception (12)	It can accurately identify the direct meaning of social cues; when multiple social cues coexist, it can accurately identify the core social cues; it can synchronously recognize and integrate multi-modal social cues; when emotional cues are present, it can accurately locate and recognize emotional information.	14%	11	Adapted (10) Self-developed (2)
	1-2 Social Understanding & Insight (12)	It can understand and infer the unexpressed true needs, thoughts, and emotional states of individuals based on external social cues (belief, intention, knowledge, emotional state); it can understand and infer the unexpressed situation-specific characteristics and needs of individuals in the current context based on external situational cues.	22%	12	Adapted (10) Self-developed (2)
2 Social Interaction	2-1 Communication (12)	It can adopt expression styles appropriate for the current context, integrating verbal and non-verbal skills to achieve natural and efficient communication outcomes; it can convey its own personality, thoughts, and preferences during communication; it can maintain emotional synchrony and interaction rhythm alignment with the communication counterpart; it can sustain multi-turn continuous dialogue while maintaining cross-turn consistency.	16%	10	Adapted (7) Self-developed (5)
	2-2 Emotional Utilization (12)	It can conduct reasoning based on emotional information during interaction, such as inferring the underlying causes or formulating coping strategies; it can adopt appropriate interaction modes or behavioral responses integrated with emotional information.	11%	10	Adapted (12) Self-developed (0)
	2-3 Relationship Management (13)	It can appropriately respond to others' thoughts and emotions during social interaction, and promote relational closeness and trust with others; it can influence and alter others' thoughts, emotions, and behaviors through non-coercive means.	5%	10	Adapted (10) Self-developed (3)
	2-4 Self-consistency (12)	In single or multiple complete interaction sessions, it can maintain the stability and integrity of its role, personality, and expression style, reflecting the stable traits of a single model.	5%	10	Adapted (10) Self-developed (2)
3 Social Learning	3-1 Observational Imitation (12)	It can learn human behavior patterns and social role expectations through observation and imitation; it can further predict others' behaviors to adjust its own responses.	8%	5	Adapted (3) Self-developed (9)
	3-2 Adaptive Learning (12)	It can learn from behavioral feedback and modify its own behaviors; it can learn the characteristics of others' short-term and long-term response patterns and generate targeted behavioral responses.	8%	6	Adapted (3) Self-developed (9)
4 Social Norm	4-1 Sociocultural Intelligence (12)	It possesses knowledge related to social interaction rules under different social and cultural contexts, including knowledge of explicit and implicit social rules, ethnic cultural knowledge, and national cultural knowledge; it can select behavioral responses appropriate for the specific sociocultural context based on the above knowledge.	4%	12	Adapted (3) Self-developed (9)
	4-2 Social Responsibility (12)	It can conduct explicit presentation of its thinking process to help humans understand the decision-making process and behavioral logic of the LLMs; it can identify and safeguard private and confidential information, including but not limited to individual privacy, enterprise trade secrets, and state secrets.	4%	9	Adapted (8) Self-developed (4)
	4-3 Moral & Ethical Intelligence (16)	It possesses knowledge related to moral and ethical norms, and can conduct moral and ethical perception, judgment, and behavioral decision-making based on the above knowledge.	3%	13	Adapted (15) Self-developed (1)
Sum	137 items	–	100%	108	Adapted (88) Self-developed (49)

Table A1: Full categories, dimensions, facets and item composition of the NICE benchmark.

B Framework Construction: Panelist Recruitment, Stage-Level Participation, and Quality Control

B.1 Panelist Recruitment and Eligibility Criteria

We recruited experts through a purposive sampling strategy designed to support interdisciplinary framework construction for social intelligence evaluation. Rather than relying on individual demographic characteristics, recruitment focused on disciplinary relevance, methodological familiarity, and task-specific expertise. Eligible panelists were required to satisfy all of the following criteria: (1) formal graduate-level training in psychology, social cognition, psychometrics, sociology, human–computer interaction, artificial intelligence, or a closely related field; (2) at least three years of research or practical experience in one of these areas; and (3) hands-on experience with at least three mainstream AI tools. These criteria were intended to ensure that panelists had sufficient domain knowledge, methodological awareness, and practical familiarity with contemporary AI systems.

B.2 Stage-Level Participation in Framework Construction

Across the framework-construction process, 23 experts contributed to one or more stages, yielding 40 stage-level participations in total. Their participation covered semi-structured interviews, structured ratings, focus-group discussion, and AHP pairwise comparisons. Because some panelists contributed to multiple stages, we report stage-level participation (Table B1).

B.3 Quality Control in Framework Construction

To ensure that panel input was used as structured methodological evidence rather than as an appeal to authority, each stage followed a predefined procedure. The interview stage used semi-structured protocols to support systematic conceptual elicitation. The structured-rating stage required panelists to evaluate candidate components using explicit criteria of necessity, relevance, and discriminability. Components showing disagreement or conceptual overlap were further discussed in the focus group, where disputed constructs were merged, removed, or redefined through structured discussion and anonymous voting. Finally, AHP pairwise comparisons were used to estimate the relative importance of the framework components.

C Benchmark Item Generation, Validation, and Quality Assessment

C.1 Item Development Team and Evaluator Eligibility

The NICE benchmark was developed and validated by researchers and evaluators with relevant interdisciplinary backgrounds. Item developers and evaluators were selected according to eligibility criteria similar to those used in framework construction, with additional emphasis on task-specific experience in item writing, social-scenario analysis, LLM evaluation, or benchmark annotation. Eligible researchers and evaluators were required to satisfy all of the following criteria: (1) formal graduate-level training in psychology, social cognition, psychometrics, sociology, human–computer interaction, artificial intelligence, or a closely related field; (2) at least three years of research or practical experience in one of these areas; and (3) hands-on experience with at least three mainstream AI tools. These requirements were intended to ensure that the benchmark development and validation process incorporated both social-scientific construct awareness and practical familiarity with LLM behavior.

C.2 Item Generation

Two researchers with backgrounds in psychology independently conducted the initial item generation and development. They referred to existing LLM evaluation datasets and classical psychological tests as development references, and then adapted or self-developed items according to the definitions and facets of each dimension in the social intelligence framework.

Specifically, one researcher was responsible for developing the social scenarios and item stems, while the other researcher designed the candidate response sets. Each item was constructed to target a specific dimension and its corresponding capability facet. To support boundary-sensitive diagnosis, each candidate response set was designed to form a quality gradient among optimal, suboptimal, and boundary-violating responses. We also ensured that each task included at least three distinct social scenarios whenever applicable.

Stage	Number of participants	Main output
Semi-structured interviews	10	Initial set of dimensions and candidate constructs
Structured ratings	16	Refined framework components and quantitative validity evidence
Focus-group discussion	5	Revised construct boundaries and operational definitions
AHP pairwise comparisons	9	Structural weights for categories and dimensions
Total	40	—

Table B1: Stage-level participation in the framework construction.

C.3 Internal Evaluation and Revision

After the initial item pool was constructed, two new internal evaluators with interdisciplinary backgrounds in computer science and artificial intelligence independently evaluated each item. Their evaluation included two components. First, they judged the optimal and worst responses for each item to examine whether the intended ranking was interpretable and stable. Second, they provided 5-point ratings for dimensional measurement validity, reliability, and neutrality.

The item developers revised the items based on the internal evaluation results and qualitative feedback. For items that received consistent ratings from both internal evaluators, the ratings were retained. For items where both evaluators agreed that any indicator—validity, reliability, or neutrality—fell below 3.5 out of 5, the item was revised and submitted for re-evaluation. For items with rating discrepancies or unstable rankings, the evaluators and item developers conducted consistency discussions until a consensus was reached or the item was substantially revised.

C.4 External Evaluation and Validation

In the final validation stage, we invited 10 additional evaluators with relevant expertise to conduct an expert evaluation of the revised item pool. These evaluators independently rated each item on dimensional measurement validity, reliability, and neutrality using the same 5-point rating scheme. Their qualitative comments were also collected to identify residual ambiguity, construct drift, cultural bias, or insufficient distinction among candidate responses.

Based on the evaluation and validation results, the development team screened, revised, and replaced items that did not meet the quality requirements. Items were retained in the final benchmark only if they satisfied the predefined quality threshold and showed acceptable expert agreement. After this iterative validation process, the final NICE benchmark contained 137 ranking items.

C.5 Evaluation Criteria and Final Scores

The final quality assessment focused on three criteria: dimensional measurement validity, reliability, and neutrality. Dimensional measurement validity refers to whether an item accurately measures the intended dimension and capability facet. Reliability refers to whether the item wording is clear, the response ranking is stable, and the gold ranking is interpretable. Neutrality refers to whether the item avoids unnecessary ambiguity, bias, or unsupported value assumptions. The final scoring results of these three indicators are reported in Table C1.

D Prompts and Model Configurations

Model Snapshots and Environment All evaluation requests were executed on April 29, 2026 (CST, UTC+8). To ensure reproducible environments, we locked our evaluation pipeline to the specific first-party API snapshots listed below:

- GPT-5.5: `gpt-5.5-20260423`
- Claude-Opus-4.7: `claude-4.7-opus-20260416`
- Gemini-3.1-pro-preview: `gemini-3.1-pro-preview-20260219`
- DeepSeek-v4-pro: `deepseek-v4-pro-20260423`
- Qwen3.6-plus: `qwen3.6-plus-04-02`

Prompt Templates Each item was presented as a two-message exchange consisting of a system instruction and a user message, authored and delivered in Simplified Chinese to match the benchmark items. English translations are shown below.

System Message

Given the situation and the question, rank the three candidate responses from

Dimension	Dimensional Measurement Validity		Reliability		Neutrality	
	M (SD)	**Avg. CV per Item**	M (SD)	**Avg. CV per Item**	M (SD)	**Avg. CV per Item**
1-Social Perception	4.4 (0.35)	17%	4.3 (0.35)	17%	4.3 (0.22)	17%
2-Social Understanding & Insight	4.4 (0.17)	15%	4.4 (0.28)	15%	4.5 (0.22)	12%
3-Communication	4.2 (0.53)	22%	4.0 (0.47)	22%	4.2 (0.35)	21%
4-Emotion Utilization	4.2 (0.29)	22%	4.1 (0.29)	21%	4.4 (0.25)	15%
5-Relationship Management	4.4 (0.24)	18%	4.3 (0.23)	19%	4.4 (0.18)	15%
6-Self-consistency	4.6 (0.13)	12%	4.5 (0.15)	14%	4.5 (0.16)	14%
7-Observational Imitation	3.8 (0.47)	23%	4.4 (0.32)	17%	4.7 (0.23)	12%
8-Adaptive Learning	4.0 (0.40)	25%	4.4 (0.24)	14%	4.6 (0.17)	13%
9-Sociocultural Intelligence	4.3 (0.29)	18%	4.3 (0.36)	17%	4.5 (0.21)	15%
10-Social Responsibility	4.3 (0.31)	15%	4.4 (0.34)	14%	4.5 (0.28)	14%
11-Moral & Ethical Intelligence	4.4 (0.33)	17%	4.3 (0.39)	17%	4.5 (0.19)	15%

Table C1: Mean (M), standard deviation (SD), and average coefficient of variation (CV) of final scores for validity, reliability, and neutrality across NICE dimensions.

best to worst. Output only the ranking in the form `2-1-3`*, with no explanation or other text.*

User Message Template

Situation: {Context}
Question: {Question}
Below are three candidate responses. Please rank them from best to second-best to worst:
1: {Opt1}
2: {Opt2}
3: {Opt3}
Output only one ranking, e.g., `2-1-3`*. Do not output any explanation.*