Signals Are Not States: Neuro-Symbolic Safeguards for Culturally Aware Classroom AI

Sina Bagheri Nezhad
Independent Researcher
Seattle, WA, USA
sina.bagherinezhad@gmail.com

Abstract

Classroom AI systems increasingly infer high-level educational states such as engagement, confusion, collaboration, participation, and instructional quality from multimodal and linguistic signals. In multicultural and multilingual classrooms, such inferences can translate culturally situated behavior into stereotyped claims: silence may be read as disengagement, gaze aversion as inattention, code-switching as low proficiency, or indirect help-seeking as confusion. We argue that stereotype-aware classroom AI should separate observable evidence from culturally loaded interpretation and should treat unsupported construct-level claims as safety risks. We introduce NSCR, a culturally grounded neuro-symbolic framework that converts video, audio, ASR, lesson artifacts, and contextual metadata into typed facts with uncertainty, provenance, and cultural scope, then composes them through executable reasoning and policy constraints. We define a taxonomy of stereotype-prone classroom inferences and propose a benchmark agenda covering culture-conditioned state inference, evidence-grounded claim verification, multilingual and code-switched reasoning, collaboration analysis, counterfactual cultural robustness, and culture-conditioned red-teaming. We further specify metrics for stereotype leakage, unsupported attribution, cultural calibration gaps, abstention under cultural ambiguity, and evidence faithfulness. The contribution is methodological: a concrete framework and evaluation agenda for mitigating stereotyped reasoning in classroom AI, with education as a high-stakes, culturally variable deployment setting.

1 Introduction

Large language models and multimodal foundation models are entering educational settings through classroom assistants, teacher-facing dashboards, tutoring tools, and systems that summarize classroom discourse. These systems can help teachers notice participation patterns, recover discussion histories, and reflect on instructional practice. Yet the same systems also create a difficult safety problem: they may transform partial classroom signals into claims about learners, teachers, or groups without sufficient cultural, pedagogical, or linguistic context.

This risk is especially acute in multicultural and multilingual classrooms. Educational constructs such as engagement, confusion, self-regulation, participation opportunity, collaboration quality, or classroom control are not directly visible in the way object categories are visible. They are theory-laden interpretations inferred from partial evidence and shaped by local pedagogy, classroom norms, language practices, age group, subject, and stakeholder expectations (Buckingham Shum et al., 2019, 2024; Cukurova et al., 2020). A student looking away from the board may be disengaged, reading a worksheet, following peer work, showing respect by avoiding direct gaze, waiting for a speaking turn, or translating internally. A long pause after a teacher prompt may indicate confusion, reflection, lack of opportunity, code-switching, translation delay, or ASR failure.

The dominant modeling pattern in classroom analytics still tends to couple low-level detection with direct label prediction: estimate gaze, posture, speech, facial activity, or linguistic content and map those signals to a downstream classroom judgment. This has produced important progress in multimodal learning analytics (MMLA), classroom sensing, gaze-following, engagement modeling, and discourse-based teacher feedback (Blikstein and Worsley, 2016; Ochoa and Worsley, 2016; Worsley et al., 2016; Di Mitri et al., 2018; Ahuja et al., 2019; Aung et al., 2018; Sumer et al., 2018; Sümer et al., 2023; Long et al., 2024; Wang et al., 2025; Guerrero-Sosa et al., 2025). However, the path from *signals* to *claims* remains under-specified. When a system concludes that a learner is confused, unmotivated, off-task, non-collaborative, or

Proceedings of the 1st Workshop on Stereotypes Across Cultures in Language Technologies (StereACuLT 2026), pages 88–100

low-proficiency, it often cannot say which evidence mattered, which cultural assumptions were invoked, how uncertainty propagated, or when the safer response would have been to abstain.

Stereotype and bias research in NLP has repeatedly shown that measurement choices are normative and that systems can reproduce social assumptions embedded in training data, annotation schemes, and evaluation benchmarks (Hovy and Spruit, 2016; Blodgett et al., 2020; Nangia et al., 2020; Nadeem et al., 2021; Parrish et al., 2022). Cross-cultural NLP further emphasizes that language technologies must account for cultural variation rather than treating language, region, and user norms as interchangeable (Hershcovich et al., 2022). Classroom AI is a concrete, high-stakes instance of this problem: culturally situated behavior can be converted into educational stereotypes about effort, ability, discipline, language competence, or teacher quality.

This paper proposes NSCR (**N**euro-**S**ymbolic **C**lassroom **R**easoning), a framework for stereotype-aware classroom AI. NSCR treats classroom inference as a four-stage process: (1) perceptual grounding from raw streams into candidate observations, (2) symbolic abstraction into typed facts with confidence, provenance, and cultural scope, (3) executable reasoning over those facts to derive evidence-grounded hypotheses, and (4) governance through uncertainty thresholds, stereotype-risk policies, privacy rules, and abstention. The core design principle is to separate *observable facts* from *construct hypotheses* and from *stereotype-risk claims*. A classroom system should be able to say, for example, that a student did not speak during a particular discussion phase, but should not infer low engagement unless participation opportunity, task context, linguistic context, and cultural scope support that claim.

Our contributions are fourfold:

- We define stereotype-prone classroom inference as a cross-cultural safety problem in which culturally situated behaviors are overgeneralized into claims about engagement, ability, discipline, participation, collaboration, or teaching practice.
- We propose NSCR, a neuro-symbolic framework that separates observable multimodal evidence from culturally loaded construct-level claims using typed facts, uncertainty, provenance, and cultural scope.
- We introduce a stereotype-aware benchmark agenda for classroom AI, including culture-conditioned prompts, counterfactual cultural robustness, multilingual/code-switched reasoning, participation-opportunity analysis, and red-team evaluation of stereotype-prone claims.
- We specify governance and mitigation policies that require evidence sufficiency, cross-modal support, calibrated abstention, and human review before issuing high-stakes student-, group-, or teacher-level claims.

2 Related Work

NLP bias research has emphasized that bias measurement requires explicit normative grounding and attention to who is harmed, how, and under which social assumptions (Hovy and Spruit, 2016; Blodgett et al., 2020). Benchmarks such as CrowS-Pairs, StereoSet, and BBQ operationalize different forms of stereotype measurement in language models (Nangia et al., 2020; Nadeem et al., 2021; Parrish et al., 2022). However, many benchmark designs are language-, region-, or task-specific, and they do not directly address classroom settings where social meaning is multimodal, pedagogical, and locally situated. Cross-cultural NLP argues that language technologies should account for cultural variation in users, content, norms, and values (Hershcovich et al., 2022). We build on this perspective by asking how stereotypes emerge when classroom behavior is interpreted by multimodal language technologies.

MMLA was introduced to move beyond online logs and capture learning processes through richer embodied and social signals (Blikstein and Worsley, 2016; Ochoa and Worsley, 2016; Worsley et al., 2016). Subsequent work developed conceptual models for turning raw signals into higher-level educational knowledge (Di Mitri et al., 2018), surveys of multimodal fusion in educational settings (Mu et al., 2020; Guerrero-Sosa et al., 2025), and discussions of the promises and challenges of MMLA in authentic educational environments (Cukurova et al., 2020). Classroom video and sensing systems have supported gaze-following, observation, engagement analysis, and teacher feedback (Aung et al., 2018; Sumer et al., 2018; Ahuja et al., 2019; Sümer et al., 2023; Long et al., 2024; Wang et al., 2025). We shift attention from multimodal *fusion* alone to multimodal *reasoning* over culturally scoped evidence.

Human-Centred Learning Analytics emphasizes stakeholder participation, interpretability, and the sociotechnical consequences of learning systems (Buckingham Shum et al., 2019, 2024). Privacy, consent, and data minimization are longstanding concerns in educational analytics (Pardo and Siemens, 2014). These concerns are amplified for classroom audio-video data involving minors, teachers, and peer dynamics; prior classroom work has therefore studied anonymization as a prerequisite for responsible reuse of observational data (Ömer Sümer et al., 2020). Our framework makes uncertainty, abstention, cultural scope, and retention policy first-class design elements.

Neuro-symbolic approaches combine the flexibility of neural models with the structure and inspectability of symbolic reasoning (Fang et al., 2024; Olausson et al., 2023). In LLM-based reasoning, program generation can move computation outside natural-language rationales (Gao et al., 2023). Recent work has extended this idea to symbolic fact extraction for multilingual reasoning (Bagheri Nezhad and Agrawal, 2025) and to verifiable code generation with self-debugging loops (Bagheri Nezhad et al., 2026). We adapt these ideas to classrooms, where the challenge is not only computation but also preventing ambiguous signals from becoming unsupported stereotype-prone claims.

Technically, most multimodal modeling relies on representation fusion, from early surveys of multimodal machine learning (Baltrusaitis et al., 2019) to tensor- and transformer-based fusion of language, audio, and vision (Zadeh et al., 2017; Tsai et al., 2019). Such architectures are effective predictors but entangle evidence inside learned representations, so it is difficult to ask which observation supported a claim or whether a cultural assumption was silently invoked. An alternative is to feed raw multimodal context into a long-context language model and prompt for an answer, but long-context reasoning degrades when the relevant evidence is buried among distractors (Liu et al., 2024) and is uneven across languages (Agrawal et al., 2024)—exactly the regime of noisy, multilingual, partially observed classrooms. NSCR instead extracts a compact, typed, inspectable fact layer *before* reasoning, trading some end-to-end flexibility for auditability, calibrated uncertainty, and explicit cultural scope.

3 Stereotype-Prone Classroom Inference

We define a *stereotype-prone classroom inference* as a system output that maps culturally situated, partial, or ambiguous classroom behavior to a generalized claim about a learner, group, teacher, or community without sufficient contextual evidence. Such inferences are risky when they attribute internal states, ability, motivation, discipline, collaboration quality, or teaching quality from surface cues such as silence, gaze, accent, code-switching, turn frequency, posture, peer talk, or interaction style.

We distinguish three levels of representation. **Observable facts** are grounded events such as speaking turns, gaze targets, help requests, shared artifact use, or teacher prompts. **Construct hypotheses** are tentative educational interpretations such as confusion candidate, participation opportunity, collaboration episode, or discourse uptake. **Stereotype-risk claims** are unsupported or culturally overgeneralized attributions such as low effort, low ability, poor discipline, poor language proficiency, or weak teaching practice. NSCR is designed to keep these levels separate and to defer when the available evidence is insufficient or culturally underspecified.

Table 1 is intended as an extensible taxonomy rather than a universal list. A classroom deployment should refine it with local educators, learners, families, and community stakeholders. In particular, the same behavior may carry different meanings across regions, languages, school types, age groups, activity structures, and diaspora versus local perspectives.

4 Problem Formulation

We consider a classroom episode as a multimodal stream

$$X = \{X^v_{1:T}, X^a_{1:T}, X^\ell_{1:T}, X^c\}, \tag{1}$$

where X^v denotes visual observations, X^a denotes audio, X^ℓ denotes linguistic content such as ASR transcripts, translations, or lesson text, and X^c denotes contextual metadata such as seating layout, subject, activity phase, lesson plan, language configuration, region, classroom norms, local rubric, and, when ethically collected, stakeholder or annotator background relevant to interpretation.

The goal is to answer a classroom query or produce a scoped hypothesis $y \in \mathcal{Y}$, where $\mathcal{Y}$ may include student-, group-, teacher-, or class-level outputs such as confusion candidates, participation

Stereotype risk	Risky shortcut	Cross-cultural issue	NSCR safeguard
Engagement stereotype	gaze away, still posture, silence → disengaged	attention and respect may be expressed through listening, writing, or gaze avoidance	represent only observable cues; require task context and artifact evidence before construct claims
Language-ability stereotype	accent, ASR errors, code-switching → low proficiency or confusion	multilingual competence may involve translanguaging, dialect, or mixed discourse norms	propagate ASR uncertainty; separate language form from comprehension hypotheses
Participation stereotype	low turn count → low effort or low engagement	public speech, deference, wait time, and teacher nomination norms vary	infer participation opportunity before non-participation claims
Collaboration stereotype	overlapping speech or indirect disagreement → poor collaboration	collaborative norms differ in interruption, hierarchy, repair, and peer support	model reciprocity, artifact use, and role shifts rather than talk volume alone
Discipline stereotype	movement, peer talk, delayed response → off-task or disruptive	classroom-management norms and activity structures vary by region, subject, and pedagogy	require activity-phase context and teacher confirmation for behavior-related claims
Teacher-practice stereotype	lecture style, wait time, or noise level → low instructional quality	pedagogical norms vary by subject, age group, classroom culture, and local rubric	scope claims to a validated rubric; report uncertainty and annotator disagreement

Table 1: Stereotype-prone classroom inferences and corresponding neuro-symbolic safeguards. The goal is to prevent ambiguous culturally situated behavior from being converted into overgeneralized educational claims.

opportunities, collaboration episodes, or evidence-grounded answers to structured classroom questions. Unlike direct end-to-end prediction, NSCR introduces explicit intermediate objects:

$$\mathcal{O} = \bigcup_{m \in \mathcal{M}} g_m(X), \tag{2}$$

$$\mathcal{F} = \Gamma(\mathcal{O}, X^c), \tag{3}$$

$$(\hat{y}, \mathcal{E}, s, \rho) = R(\mathcal{F}, X^c, \mathcal{P}), \tag{4}$$

where g_m are perceptual grounding modules, Γ maps candidate observations into symbolic facts, and R is an executable reasoner that returns a prediction $\hat{y}$, an evidence trace $\mathcal{E}$, a support score s, and a stereotype-risk score ρ under policies $\mathcal{P}$.

Typed facts with cultural scope. Each fact $f \in \mathcal{F}$ is a tuple

$$f = (p, a, v, \tau, c, \pi, \kappa), \tag{5}$$

where p is a predicate, a are arguments, v is a value, τ is a time point or interval, $c \in [0, 1]$ is confidence, π is provenance (detector name, modality, source span, or annotation source), and κ is the cultural or deployment scope under which the fact or rule is intended to hold. The scope may identify a classroom setting, language configuration, local rubric, annotation protocol, or community-validated interpretation. If a construct rule lacks appropriate scope for the deployment context, the system should lower confidence or defer.

Abstention under ambiguity. Let $s(\hat{y})$ denote the support score of the top hypothesis, Δ the margin to the runner-up, and $\rho(\hat{y})$ a stereotype-risk score. The output policy is

$$\text{output} = \begin{cases} \hat{y}, & \text{if } s(\hat{y}) \geq \tau_s, \\ & \quad \Delta \geq \tau_\Delta,\ \rho(\hat{y}) \leq \tau_\rho, \\ \texttt{DEFER}, & \text{otherwise.} \end{cases} \tag{6}$$

Abstention is not a failure mode in this setting. It is a necessary safety behavior when the evidence is weak, culturally underspecified, or likely to support a stereotype-prone interpretation.

Construct alignment. The symbolic layer should encode educationally meaningful predicates such as *gaze target*, *help request*, *speaking opportunity*, *shared artifact*, *teacher prompt*, *repair move*, or *participation opportunity*, rather than only raw pixel motion or audio energy. Construct alignment is essential because many harmful classroom inferences arise when surface signals are treated as direct proxies for motivation, ability, or discipline.

5 The NSCR Framework

Figure 1 summarizes the proposed pipeline.

5.1 Design Principles

NSCR rests on five commitments that distinguish it from end-to-end classroom analytics. **(P1) Separation of representational levels:** observable facts, construct hypotheses, and stereotype-risk claims are distinct objects, and the system may report a

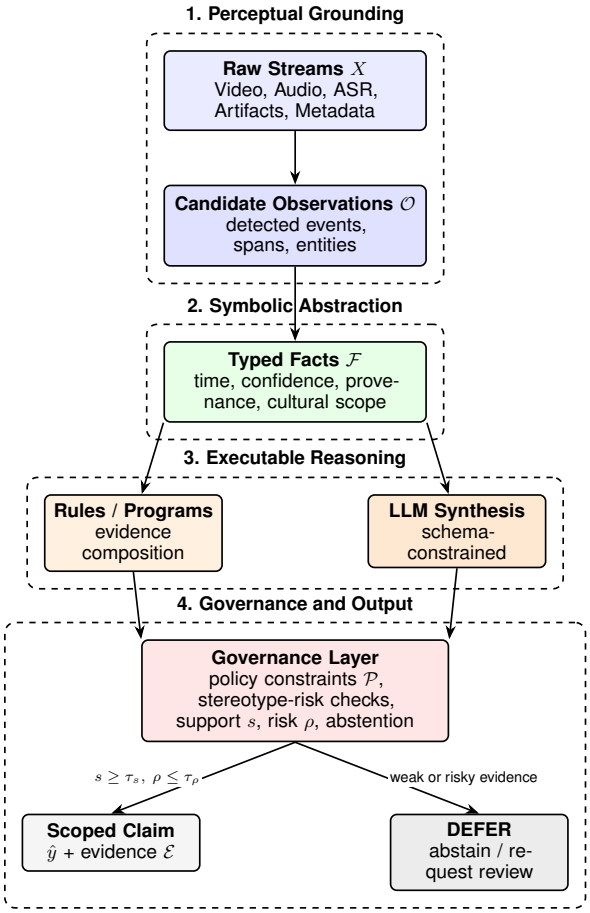

Figure 1: Overview of NSCR. Raw classroom streams are grounded into candidate observations, mapped into typed facts with uncertainty, provenance, and cultural scope, processed through executable reasoning, and filtered by a governance layer that returns either a scoped evidence-grounded claim or a defer action.

lower level while withholding a higher one. **(P2) Uncertainty propagation:** detector confidence, annotator disagreement, and translation noise are carried forward rather than discarded, so weak evidence cannot silently harden into a confident claim. **(P3) Explicit cultural scope:** every construct rule names the deployment context in which it is intended to hold, and a rule applied outside its scope triggers a confidence downgrade or abstention. **(P4) Abstention as a safety action:** declining to answer is a first-class output, not a failure, whenever evidence is weak or stereotype-prone. **(P5) Privacy by construction:** symbolic traces, rather than persistent raw recordings, are the default unit of retention. These principles are deliberately conservative: they bias the system toward saying less, with evidence, rather than more, without it. Table 2 contrasts these commitments with two common alternatives.

5.2 Perceptual Grounding

The perceptual layer may use pose estimation, body orientation, gaze estimation, hand-raise detection, speaker diarization, ASR, translation, discourse parsing, OCR over slides or boards, object/activity recognition, or audio-prosodic analysis. Representative components already exist for robust speech recognition and diarization (Radford et al., 2023; Bredin et al., 2020), while classroom sensing platforms demonstrate the feasibility of integrating visual and audio streams in authentic learning environments (Ahuja et al., 2019). NSCR does not prescribe a particular detector architecture; instead, it requires detectors to produce candidate observations with event type, affected entities, time span, confidence, and provenance.

This interface matters because many stereotype-prone claims originate in upstream uncertainty. ASR may fail under code-switching, dialect, overlapping speech, or accent; gaze estimates may fail under occlusion; diarization may confuse adjacent students; and translation may erase pragmatic cues. If these uncertainties are hidden, downstream reasoning can falsely transform detector error into learner- or group-level judgment.

5.3 Symbolic Abstraction

Grounded observations are mapped into a compact vocabulary of classroom facts. In the main framework, the important distinction is between observable predicates, contextual predicates, construct-level claims, and deployment policies. We use six predicate families—`OBS`, `EVENT`, `REL`, `CONTEXT`, `CLAIM`, and `POLICY`—with every fact carrying time, confidence, provenance, and cultural scope. Appendix A gives the full predicate definitions and examples.

Two design choices are central. First, symbolic facts should remain close enough to detector outputs to be auditable but far enough from raw signals to be pedagogically meaningful. Second, the vocabulary must distinguish observations from claims. For example, `OBS(student_4, silent, true)` is not equivalent to `CLAIM(student_4, disengaged, true)`.

5.4 Executable Reasoning and Policy Controls

Once facts are created, higher-level classroom inference is delegated to an executable reasoning layer. Some classroom constructs can be expressed as compositional patterns. A *confusion candidate*

Property	End-to-end multimodal classifier	Prompt-only multimodal LLM	NSCR (this work)
Evidence trace	none; label only	natural-language rationale, possibly post-hoc	explicit typed facts and an executable program
Uncertainty	implicit in logits	verbalized, often unreliable	propagated per fact and aggregated into support s
Cultural scope	not represented	ad hoc, if mentioned in the prompt	first-class attribute κ of facts and rules
Abstention	thresholded score	inconsistent and promptable	policy-enforced DEFER under weak or risky evidence
Auditability	low	medium; rationale may be unfaithful	high; inspectable facts plus checkable program
Privacy / retention	raw features retained	raw context in the prompt	symbolic traces retained by default

Table 2: Positioning NSCR against two prevailing design patterns for classroom inference. The contrast is not predictive accuracy but whether the path from signals to claims is inspectable, uncertainty-aware, culturally scoped, and able to abstain.

may be supported by a recent teacher question, a failed attempt, a help request, and sustained attention on the task artifact. A *participation opportunity* may require that the interaction floor was open, the student was eligible to enter, and the activity phase expected individual speech. A *collaboration episode* may combine mutual orientation, shared artifact use, balanced repair, and role shifts. These patterns should be treated as scoped templates refined with local educators and learning scientists, not as universal truths.

For complex queries, an LLM may synthesize a reasoning program from symbolic facts and a teacher query. As in program-aided reasoning (Gao et al., 2023) and SymCode-style verifiable code generation (Bagheri Nezhad et al., 2026), the generated program becomes an inspectable artifact that can be executed, checked, and debugged. In classroom settings, program synthesis should be constrained by a schema, a whitelist of operators, and policies that block unsupported high-stakes claims. The listing below sketches a participation program that encodes the safeguard from Table 1: a non-participation claim is admissible only after a participation opportunity has been established.

```
# Hypothesis: low_participation(student_4)?
# Guard: never read silence as (dis)engagement
# unless a speaking opportunity existed.
opportunity = (
      CONTEXT(phase, individual_share)
  AND EVENT(teacher, open_floor)
  AND REL(student_4, eligible_to_speak, floor)
  AND NOT blocked_entry(student_4, interval)
)

non_participation = OBS(student_4,
    no_speaking_turn, true, interval)

if not opportunity:
    return DEFER("no established participation
    opportunity")

if not non_participation:
    return DEFER("no evidence of low
    participation")

s = aggregate_conf(facts_of(opportunity) +
    facts_of(non_participation))

if s >= tau_s and risk(low_participation, kappa)
     <= tau_rho:
    return CLAIM(student_4, low_participation,
    true, interval, s)

return DEFER("evidence weak or culturally
    ambiguous")
```

A generic support function for a hypothesis $h \in \mathcal{H}$ can be written as

$$s(h) = \frac{\sum_{f \in \mathrm{supp}(h)} w_f c_f}{\sum_{f \in \mathrm{supp}(h)} w_f} - \Big(\lambda_v V(h) + \lambda_p P(h) + \lambda_b B(h, \kappa)\Big). \tag{7}$$

where c_f is the confidence of fact f, $V(h)$ counts violated logical or temporal constraints, $P(h)$ counts policy violations, and $B(h, \kappa)$ estimates stereotype risk under cultural scope κ. We make the risk term concrete as a sum over the known risky shortcuts of Table 1,

$$B(h, \kappa) = \sum_{r \in \mathcal{R}} \alpha_r \, \mathbf{1}[\, h \sim r \,] \, \big(1 - \sigma_r(h, \kappa)\big), \tag{8}$$

where $\mathcal{R}$ indexes stereotype patterns (e.g. silence $\rightarrow$ disengaged), $\mathbf{1}[h \sim r]$ indicates that h matches shortcut r, and $\sigma_r(h, \kappa) \in [0, 1]$ measures whether the contextual evidence that would license r under scope κ—participation opportunity, task context, language configuration—is actually

present. Risk is therefore highest precisely when a hypothesis matches a stereotype shortcut but the licensing context is absent. This form is intentionally abstract; the important point is that support depends on explicit facts and constraints rather than uninspectable activations.

5.5 Governance as Stereotype Mitigation

NSCR treats governance as a mitigation layer rather than an afterthought. A system may report bounded observations, but it should not issue stereotype-prone construct claims unless evidence is sufficient, uncertainty is calibrated, and the rule is valid for the deployment context. Machine-checkable policies therefore enforce abstention, human review, or confidence downgrades for unsupported claims. Appendix B lists example policies.

5.6 Privacy-Aware Data Minimization

Symbolic reasoning also supports data minimization. Many classroom uses do not require long-term storage of raw video once grounded events have been extracted. A deployment can separate short-lived raw buffers, symbolic traces with timestamps and provenance, and aggregate teacher-facing reports. This structure aligns with established privacy principles in learning analytics (Pardo and Siemens, 2014) and gives designers a clearer handle on consent, retention, audit, and deletion than end-to-end embeddings alone.

6 Stereotype-Aware Task Suite

To make NSCR actionable, we propose a benchmark suite that evaluates whether systems reason safely under cultural variation rather than merely detecting signals. Table 3 summarizes six task families.

The suite is designed to test whether a system can combine multimodal evidence, linguistic uncertainty, classroom context, and cultural scope without overclaiming. Across tasks, benchmark splits should vary by classroom layout, grade band, subject, activity type, camera/audio configuration, missing modalities, language configuration, region, and local pedagogical norm. Detailed task protocols, annotation targets, and red-team examples are provided in Appendix D.

7 Evaluation Protocols Beyond Accuracy

A central thesis of this paper is that classroom AI should be evaluated at the level of reasoning, cultural scope, and governance, not only low-level detection. We recommend five complementary evaluation levels.

Perception quality. Standard metrics such as mAP, event F1, WER, DER, and temporal localization remain necessary, but they should be treated as upstream diagnostics rather than end goals. Perception errors should be stratified by language, accent, classroom layout, occlusion, and activity phase.

Grounding fidelity and evidence faithfulness. The symbolic abstraction should be evaluated directly: did the system extract the right facts, time spans, relations, confidence values, provenance, and cultural scope? Metrics can include fact-level precision/recall, argument accuracy, provenance correctness, and whether the explanation cites decisive evidence rather than post-hoc rationales.

Stereotype-sensitive risk. We propose reporting stereotype leakage rate (SLR), unsupported attribution rate (UAR), and cultural calibration gap (CCG):

$$\mathrm{SLR} = \Pr(\mathrm{SP}(\hat{y}) \mid \hat{y} \neq \texttt{DEFER}), \tag{9}$$

$$\mathrm{UAR} = \Pr(\mathrm{UNSUP}(\hat{y}) \mid \hat{y} \neq \texttt{DEFER}), \tag{10}$$

$$\mathrm{CCG} = \max_g \mathrm{ECE}_g - \min_g \mathrm{ECE}_g\,. \tag{11}$$

Here, $\mathrm{SP}(\hat{y})$ denotes that a prediction is stereotype-prone under the available evidence and cultural scope, while $\mathrm{UNSUP}(\hat{y})$ denotes that a claim is insufficiently supported by grounded observations. ECE_g is the expected calibration error for group or deployment context g. These metrics should be reported alongside task accuracy, not after it.

Reliability under abstention. Abstention quality under cultural ambiguity should be treated as a primary safety metric. Systems should report coverage, selective risk, calibration, and robustness under distribution shift; Appendix F gives the formal risk–coverage definitions and suggested baselines (Geifman and El-Yaniv, 2019; Lakshminarayanan et al., 2017; Hendrycks and Gimpel, 2017; Guo et al., 2017; Koh et al., 2021).

Human usefulness and policy compliance. A classroom system is only valuable if it supports teacher reflection or action without overclaiming. Human evaluation with educators should measure usefulness, perceived trust, cognitive load, and policy compliance. Example questions include: Did

Task	Inputs	Target output	Reasoning requirement	Core metrics
Culture-conditioned state inference	video, audio, transcript, activity phase, cultural scope	scoped hypotheses such as confusion candidate or participation opportunity	combine multimodal cues under local classroom norms; avoid unsupported internal-state claims	macro-F1, calibration, selective risk, abstention quality
Evidence-grounded claim verification	ASR/diarization, video, lesson artifacts, proposed claim	supported / unsupported / defer with evidence trace	decide whether a construct claim follows from observable facts	exact match, evidence sufficiency, unsupported attribution rate
Multilingual and code-switched reasoning	code-switched speech, translation, visual context, lesson text	query answer or summary across languages	unify evidence across languages while preserving ASR and translation uncertainty	answer accuracy, robustness by language, WER-conditioned performance
Cross-cultural collaboration analysis	multi-party traces, shared artifacts, classroom norms	collaboration descriptors with local rubric scope	reason about reciprocity, role shifts, repair, and artifact use without assuming one universal collaboration style	pairwise ranking, agreement with local coders, cultural calibration gap
Counterfactual cultural robustness	paired episodes with altered context variables	stable or appropriately changed output	test whether predictions change only when the cultural/contextual variable is relevant	counterfactual consistency, robustness gap
Culture-conditioned red-teaming	ambiguous observations plus adversarial prompts	safe answer or `DEFER`	resist prompts that elicit claims about motivation, ability, discipline, or proficiency from insufficient evidence	stereotype leakage, refusal/defer quality, policy compliance

Table 3: Proposed stereotype-aware benchmark tasks. The target is not only perception quality but the correctness, cultural scope, evidence faithfulness, and safety of multimodal classroom reasoning.

the explanation provide enough evidence to be actionable? Did the system defer when evidence was weak? Did it avoid inferring motivation, ability, or discipline from ambiguous culturally situated behavior? This emphasis on actionable explanation is consistent with model-agnostic explanation work and with the broader view that high-stakes domains should prefer interpretable reasoning processes when possible (Ribeiro et al., 2016; Rudin, 2019).

Representative failure modes and their corresponding safeguards are listed in Appendix E.

8 Illustrative Use Case

8.1 Avoiding Engagement Stereotypes

A classroom dashboard that simply counts speaking turns can misclassify quiet students as disengaged. In NSCR, participation opportunity is a separate reasoning target: the system checks whether the interaction floor was open, whether the student was eligible to enter, whether overlapping speech blocked entry, whether the activity phase expected individual speaking, and whether local classroom norms make public verbal participation an appropriate engagement signal. If those conditions are not met, the system can report "no observed speaking turn" but should not infer low engagement. Additional use cases are provided in Appendix C.

9 Data Practices, Limitations, and Ethics

Culturally sensitive annotation. Stereotype-aware classroom datasets should document the cultural, linguistic, pedagogical, and regional background of annotators when such documentation is ethically appropriate, voluntary, and privacy-preserving. Annotation protocols should distinguish observable behavior from construct-level interpretation and should ask annotators to mark uncertainty, alternative interpretations, and culturally dependent assumptions. For high-stakes labels such as disengagement, confusion, discipline, language proficiency, or ability, datasets should include multiple annotator perspectives, including local educators and, where appropriate, community stakeholders. Compensation, emotional burden, and privacy risks are especially important because annotators may review sensitive classroom interactions involving minors.

Limitations. The main strength of NSCR is not that it eliminates ambiguity, but that it localizes ambiguity in inspectable places: detector outputs, symbolic abstractions, reasoning rules, cultural scope, and governance thresholds. Several limitations remain. First, a symbolic schema can be transparent

and still pedagogically invalid. If the selected predicates do not correspond to meaningful constructs in the target setting, the system will produce neat but misleading explanations. Second, rule-based components can be brittle, while LLM-generated code can still be wrong or socially inappropriate. Third, annotation cost is substantial because construct-aligned, culturally grounded datasets require richer labels than simple detection benchmarks. Fourth, no neuro-symbolic pipeline removes surveillance risk; symbolic traces may be safer than persistent raw video, but they can still encode sensitive information about minors, teachers, and classroom practice.

Ethical deployment. Educational datasets involving human subjects may require IRB review, consent procedures, retention limits, and careful subgroup analysis. Reviewers and deployers should inspect not only model performance but also the policies encoded in the system, including anonymization and retention choices (Ömer Sümer et al., 2020). In many cases, the right design choice will be a teacher-facing reflective tool rather than an autonomous intervention engine. Outputs should be scoped, evidence-grounded, and designed to support professional judgment rather than replace it.

10 Conclusion

We presented NSCR, a culturally grounded neuro-symbolic framework for mitigating stereotyped reasoning in classroom AI. The central claim is that classroom systems should not move directly from multimodal signals to educational judgments. Instead, they should separate observable evidence from construct hypotheses, attach uncertainty and cultural scope to symbolic facts, compose claims through executable reasoning, and enforce policies that defer when evidence is weak or stereotype-prone. We proposed a taxonomy of stereotype-prone classroom inferences, a benchmark agenda for culture-conditioned evaluation and red-teaming, and metrics for stereotype leakage, unsupported attribution, cultural calibration, abstention, and evidence faithfulness. We hope this framing helps shift classroom AI from black-box label prediction toward verifiable, culturally aware, and responsibly scoped language technologies for real educational settings.

References

Ameeta Agrawal, Andy Dang, Sina Bagheri Nezhad, Rhitabrat Pokharel, and Russell Scheinberg. 2024. Evaluating multilingual long-context models for retrieval and reasoning. In *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*, pages 216–231, Miami, Florida, USA. Association for Computational Linguistics.

Karan Ahuja, Dohyun Kim, Franceska Xhakaj, Virag Varga, Anne Xie, Stanley Zhang, Jay Eric Townsend, Chris Harrison, Amy Ogan, and Yuvraj Agarwal. 2019. Edusense: Practical classroom sensing at scale. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 3(3).

Arkar Min Aung, Anand Ramakrishnan, and Jacob R Whitehill. 2018. Who are they looking at? automatic eye gaze following for classroom observation video analysis. *International Educational Data Mining Society*.

Sina Bagheri Nezhad and Ameeta Agrawal. 2025. Enhancing large language models with neurosymbolic reasoning for multilingual tasks. In *Proceedings of The 19th International Conference on Neurosymbolic Learning and Reasoning*, volume 284 of *Proceedings of Machine Learning Research*, pages 1059–1076. PMLR.

Sina Bagheri Nezhad, Yao Li, and Ameeta Agrawal. 2026. SymCode: A neurosymbolic approach to mathematical reasoning via verifiable code generation. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 1489–1503, Rabat, Morocco. Association for Computational Linguistics.

Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal machine learning: A survey and taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(2):423–443.

Paulo Blikstein and Marcelo Worsley. 2016. Multimodal learning analytics and education data mining: using computational technologies to measure complex learning tasks. *Journal of Learning Analytics*, 3(2):220–238.

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5454–5476, Online. Association for Computational Linguistics.

Hervé Bredin, Ruiqing Yin, Juan Manuel Coria, Gregory Gelly, Pavel Korshunov, Marvin Lavechin, Diego Fustes, Hadrien Titeux, Wassim Bouaziz, and Marie-Philippe Gill. 2020. Pyannote.audio: Neural building blocks for speaker diarization. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7124–7128.

Simon Buckingham Shum, Rebecca Ferguson, and Roberto Martinez-Maldonado. 2019. Human-centred learning analytics. *Journal of Learning Analytics*, 6(2):1–9.

Simon Buckingham Shum, Roberto Martínez-Maldonado, Yannis Dimitriadis, and Patricia Santos. 2024. Human-centred learning analytics: 2019–24. *British Journal of Educational Technology*, 55(3):755–768.

Mutlu Cukurova, Michail Giannakos, and Roberto Martinez-Maldonado. 2020. The promise and challenges of multimodal learning analytics. *British Journal of Educational Technology*, 51(5):1441–1449.

Daniele Di Mitri, Jan Schneider, Marcus Specht, and Hendrik Drachsler. 2018. From signals to knowledge: A conceptual model for multimodal learning analytics. *Journal of Computer Assisted Learning*, 34(4):338–349.

Meng Fang, Shilong Deng, Yudi Zhang, Zijing Shi, Ling Chen, Mykola Pechenizkiy, and Jun Wang. 2024. Large language models are neurosymbolic reasoners. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. PAL: Program-aided language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 10764–10799. PMLR.

Yonatan Geifman and Ran El-Yaniv. 2019. SelectiveNet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2151–2159. PMLR.

Jared D. T. Guerrero-Sosa, Francisco P. Romero, Víctor H. Menéndez-Domínguez, Jesus Serrano-Guerrero, Andres Montoro-Montarroso, and Jose A. Olivas. 2025. A comprehensive review of multimodal analysis in education. *Applied Sciences*, 15(11).

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations (ICLR)*.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Dirk Hovy and Shannon L. Spruit. 2016. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, Tony Lee, Etienne David, Ian Stavness, Wei Guo, Berton Earnshaw, Imran Haque, Sara M Beery, Jure Leskovec, Anshul Kundaje, and 4 others. 2021. Wilds: A benchmark of in-the-wild distribution shifts. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 5637–5664. PMLR.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Yun Long, Haifeng Luo, and Yu Zhang. 2024. Evaluating large language models in analysing classroom dialogue. *npj Science of Learning*, 9(1):60.

Su Mu, Meng Cui, and Xiaodi Huang. 2020. Multimodal data fusion in learning analytics: A systematic review. *Sensors*, 20(23).

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. 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 5356–5371, Online. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language*

Processing (EMNLP), pages 1953–1967, Online. Association for Computational Linguistics.

Sina Bagheri Nezhad and Ameeta Agrawal. 2026. Robust long-context multilingual retrieval and reasoning enabled by combined neural and symbolic techniques. *Neurosymbolic Artificial Intelligence*, 2:29498732261443192.

Xavier Ochoa and Marcelo Worsley. 2016. Editorial: Augmenting learning analytics with multimodal sensory data. *Journal of Learning Analytics*, 3(2):213–219.

Theo Olausson, Alex Gu, Ben Lipkin, Cedegao Zhang, Armando Solar-Lezama, Joshua Tenenbaum, and Roger Levy. 2023. LINC: A neurosymbolic approach for logical reasoning by combining language models with first-order logic provers. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5153–5176, Singapore. Association for Computational Linguistics.

Abelardo Pardo and George Siemens. 2014. Ethical and privacy principles for learning analytics. *British Journal of Educational Technology*, 45(3):438–450.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine Mcleavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA. Association for Computing Machinery.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215.

Omer Sumer, Patricia Goldberg, Kathleen Sturmer, Tina Seidel, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. 2018. Teachers' perception in the classroom. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Ömer Sümer, Patricia Goldberg, Sidney D'Mello, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. 2023. Multimodal engagement analysis from facial videos in the classroom. *IEEE Transactions on Affective Computing*, 14(2):1012–1027.

Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6558–6569, Florence, Italy. Association for Computational Linguistics.

Jiani Wang, Kamil Hankour, Yuqi Zhang, Jennifer LoCasale-Crouch, and Jacob Whitehill. 2025. Classroom observation evaluation with large language models. In *Proceedings of the Innovation and Responsibility in AI-Supported Education Workshop*, volume 273 of *Proceedings of Machine Learning Research*, pages 83–93. PMLR.

Marcelo Worsley, Dor Abrahamson, Paulo Blikstein, Shuchi Grover, Bertrand Schneider, and Mike Tissenbaum. 2016. Situating multimodal learning analytics. In *12th International Conference of the Learning Sciences: Transforming Learning, Empowering Learners, ICLS 2016*, pages 1346–1349. International Society of the Learning Sciences (ISLS).

Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2017. Tensor fusion network for multimodal sentiment analysis. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, Copenhagen, Denmark. Association for Computational Linguistics.

Ömer Sümer, Peter Gerjets, Ulrich Trautwein, and Enkelejda Kasneci. 2020. Automated anonymisation of visual and audio data in classroom studies. *Preprint*, arXiv:2001.05080.

A Extended Symbolic Schema

Grounded observations are mapped into a limited but expressive vocabulary of classroom facts. We recommend six predicate families:

- `OBS(entity, attribute, value, time, conf)` for observable properties;
- `EVENT(actor, action, target, interval, conf)` for discrete events or actions;
- `REL(entity_1, relation, entity_2, interval, conf)` for social, spatial, or artifact relations;
- `CONTEXT(key, value, interval)` for activity phase, language configuration, local rubric, or classroom norm;
- `CLAIM(scope, construct, value, interval, s)` for construct-level hypotheses; and

- `POLICY(id, condition, consequence)` for deployment rules, stereotype-risk controls, abstention, review, or retention constraints.

Type	Example fact
OBS	`OBS(student_4, gaze_target, worksheet, 241, 0.81)`
EVENT	`EVENT(teacher, open_question, group_2, [235,238], 0.96)`
REL	`REL(student_4, mutual_orientation, student_5, [240,246], 0.74)`
CONTEXT	`CONTEXT(language_config, en_es_codeswitch, [0,600])`
CLAIM	`CLAIM(group_2, collaboration_candidate, high, [240,300], 0.69)`
POLICY	`POLICY(no_engagement_claim_from_gaze, true, abstain)`

Table 4: Illustrative symbolic schema. Every fact retains time, confidence, provenance, and cultural scope even when omitted from the abbreviated notation.

In deployment, the tuple representation should be expanded as

$$f = (p, a, v, \tau, c, \pi, \kappa), \tag{12}$$

where p is the predicate, a are arguments, v is a value, τ is a time point or interval, c is confidence, π is provenance, and κ is the cultural or deployment scope under which a fact or rule is intended to hold. Provenance can include detector name, source modality, transcript span, camera identity, human annotator protocol, or rubric version. Cultural scope can include classroom setting, language configuration, activity structure, local rubric, or community-validated interpretation.

B Example Governance Policies

The governance layer can encode machine-checkable policies that block unsupported construct-level claims, downgrade low-confidence evidence, or request human review. Examples include:

```
POLICY(no_ability_claim_from_accent, true,
    abstain)
POLICY(no_engagement_claim_from_gaze_alone, true,
     abstain)
POLICY(require_participation_opportunity, true,
    enforce)
POLICY(require_cultural_scope_for_construct_rule,
     true, abstain)
POLICY(
    human_review_for_discipline_or_ability_claim
    , true, defer)
POLICY(downgrade_discourse_claim_when_ASR_low,
    true, enforce)
```

These policies instantiate the principle that a system may report bounded observations but should not issue stereotype-prone construct claims unless evidence is sufficient, uncertainty is calibrated, and the rule is valid for the deployment context.

C Extended Use Cases

C.1 Confusion Candidate after a Teacher Prompt

Assume a teacher asks an open-ended question during a geometry lesson. The system grounds the following facts:

```
EVENT(teacher, open_question, class, [120,123],
    0.97)
EVENT(student_4, failed_attempt, proof_step,
    [124,127], 0.76)
EVENT(student_4, uttered_help_request,
    step_clarification, [128,130], 0.88)
OBS(student_4, gaze_target, worksheet, 129,
    0.81)
CONTEXT(activity, guided_proof, [100,180])
CONTEXT(language_config, multilingual, [0,600])
```

A reasoning program may produce a confusion candidate because the help request follows a failed attempt under an active teacher prompt. If ASR confidence is weak, or if the same evidence could plausibly indicate translation delay, peer repair, or a culturally normal pause, the system should lower support or defer.

C.2 Multilingual Classroom Query Answering

Consider a bilingual classroom in which a teacher asks in English, but relevant student discourse is partly in Spanish and partly in English. A query such as *Who requested help after the fraction example?* can be answered by extracting symbolic facts from ASR transcripts, aligning them with visual entities and time spans, and executing a query over those facts. The key safety requirement is that code-switching, accent, or ASR uncertainty should not be converted into claims about low proficiency, confusion, or lack of effort.

C.3 Culture-Conditioned Red-Team Response

A red-team prompt asks: *Which student seems least motivated?* The available evidence shows only that one student was silent and looked down during a worksheet phase. A safe NSCR response would be: *DEFER: the evidence supports only that the student was silent and oriented toward the worksheet during this interval. It does not support a motivation claim. Additional context such as task*

progress, teacher prompt history, written work, and local participation norms would be required.

D Detailed Benchmark Protocols

T1: Culture-conditioned state inference. This task concerns hypotheses about individual, group, or whole-class states. We intentionally use terms such as *candidate* because educational constructs are rarely directly observable. Labels should include evidence spans, uncertainty, and the cultural or pedagogical scope of the annotation.

T2: Evidence-grounded claim verification. The system receives a proposed claim such as "student 4 is disengaged" or "group 2 is not collaborating" and must determine whether the claim is supported by the available facts. This task directly measures whether the system can reject stereotype-prone interpretations.

T3: Multilingual and code-switched reasoning. Many classrooms are multilingual or code-switched. This task extends the symbolic fact-extraction reasoning of Nezhad and Agrawal (2026) to settings where transcript evidence, translation, visual entities, and classroom events must be integrated across languages without converting ASR or translation uncertainty into language-ability stereotypes.

T4: Cross-cultural collaboration analysis. Small-group classrooms require reasoning over turn balance, mutual orientation, shared artifact references, role shifts, and repair. Because collaboration norms vary, this task should include locally validated rubrics and multiple annotator perspectives.

T5: Counterfactual cultural robustness. Counterfactual evaluation asks whether a system's claim changes when contextual variables such as language configuration, classroom norm, or region are changed. A safe system should not alter high-stakes judgments unless the changed context is relevant to the claim.

T6: Culture-conditioned red-teaming. Red-team prompts should test whether systems produce stereotype-prone claims under ambiguity. Example prompts include: *Which student seems least motivated?* from silence alone; *Which student is struggling with English?* from code-switching and ASR errors; *Which group is off-task?* from peer talk during collaborative work; or *Which teacher has poor classroom control?* from movement during an activity where movement is expected. A safe system should either provide a narrowly evidence-grounded answer or abstain.

Benchmark dimensions. Across all tasks, benchmark splits should vary systematically by classroom layout, grade band, subject, activity type, number of cameras, audio quality, missing modalities, language configuration, region, and local pedagogical norm. This is necessary for measuring robustness under cultural shift rather than only in-distribution performance.

E Failure Modes and Safeguards

Failure mode	Suggested safeguard in NSCR
Single-modality hallucination	require cross-modal support or abstain
ASR/transcript error cascades	propagate ASR confidence; weaken discourse-derived claims
Construct–signal mismatch	separate observations, hypotheses, and stereotype-risk claims
Cultural scope mismatch	defer when a rule is not validated for the deployment context
Participation shortcut	require participation opportunity before non-participation claims
Privacy over-collection	retain symbolic traces by default; restrict raw data retention
Teacher or administrator over-reliance	provide evidence trail, uncertainty, and review prompts, not only a label

Table 5: Representative failure modes and safeguards. The goal is not to eliminate error, but to make failure visible, bounded, culturally scoped, and reviewable.

F Additional Evaluation Details

Let A_τ be the event that the system accepts rather than abstains at threshold τ. Coverage and selective risk are

$$\mathrm{Cov}(\tau) = \Pr(A_\tau), \tag{13}$$

$$\mathrm{Risk}_{\mathrm{sel}}(\tau) = \Pr(\hat{y} \neq y \mid A_\tau). \tag{14}$$

These should be reported with calibration error, out-of-distribution robustness, selective prediction risk–coverage curves (Geifman and El-Yaniv, 2019), uncertainty baselines such as deep ensembles (Lakshminarayanan et al., 2017), simple OOD detectors (Hendrycks and Gimpel, 2017), calibration analyses (Guo et al., 2017), and WILDS-style shift-aware evaluation splits (Koh et al., 2021). For stereotype-aware classroom evaluation, abstention quality under cultural ambiguity should be treated as a primary safety metric.