Confidently Wrong Across Contexts: LLM-as-Judge Reliability in Tool-or-Teammate Framing Detection

Anonymous ACL submission

Abstract

How practitioners describe workplace LLMs – as tools, teammates, mixed, or systems to reject or defer – carries consequences for accountability, oversight, and organisational acceptability. We introduce a four-class task for detecting these framings in semi-structured interviews with 35 practitioners across India and the UK, with a grounded-theory codebook (Cohen's $\kappa = 0.565$). On a held-out test set, fine-tuned RoBERTa-base approaches the human inter-rater reference (macro-F_1 = 0.547 $\pm$ 0.033), exceeding a TF-IDF baseline (0.480). Evaluating four LLM judges under an identical protocol, we find a large frontier–hosted-open gap: frontier judges (Claude Sonnet 4.5, GPT-4o) reach 0.65–0.75 macro-F_1 while open-weight judges stay below the lexical baseline at 0.39–0.44. Three failure modes follow: miscalibration that is moderate for frontier judges (ECE 0.09–0.13) but severe for open-weight judges, which are *confidently wrong* (confidence $\geq$ 0.94, high-confidence error 50–58%); cross-vendor agreement near chance ($\kappa = 0.085$) yet $\kappa = 0.879$ between the two open-weight judges, a shared failure pattern; and high sensitivity to exemplar selection. Cross-national transfer degrades encoder fine-tuning sharply while TF-IDF transfers more robustly. We document a restricted-access corpus and release the codebook, schema, prompts, prediction files, and diagnostics. Our results caution against treating LLM-as-judge protocols as interchangeable drop-in substitutes for human annotation: judge choice is a first-order measurement variable that should not be assumed without task-specific validation and transparent reporting.

1 Introduction

When workplace practitioners adopt LLMs into their workflows, they describe the model in heterogeneous ways: as a draft generator under their control, as a collaborator that talks back, as something whose role they cannot quite pin down, or as something they have explicitly decided not to use. These framings are not stylistic. They carry assumptions about accountability, oversight, and organisational acceptability, and shape downstream decisions about how a model is integrated, audited, and governed. Recent HCI scholarship has documented this framing variation through interviews and ethnographies. The corresponding NLP question, whether such framings can be detected computationally and where detection fails, has not been systematically examined. The question matters both for practitioners (it is the linguistic surface against which policies for LLM adoption are written) and for the NLP community (it is a non-trivial subjective discourse classification task where frontier LLM-as-judge protocols are already being proposed as substitutes for human coding).

We introduce *Tool-or-Teammate Framing Detection*, a four-class classification task derived from grounded-theory coding (Strauss and Corbin, 1990) of 35 semi-structured interviews with designers, developers, and product practitioners in India ($n = 16$) and the United Kingdom ($n = 19$). The four labels, *tool*, *teammate*, *mixed*, and *reject_or_defer*, capture distinct positions practitioners take when describing LLM use. Annotation follows a codebook developed through four rounds of coder calibration; each interview turn is independently coded by two annotators with adjudication, yielding Cohen's $\kappa = 0.565$ on the classifier subset.

We compare four classifier configurations: a majority-class floor; a TF-IDF baseline with class-weighted logistic regression; a fine-tuned RoBERTa-base encoder (Liu et al., 2019), evaluated as a five-seed sweep with bootstrapped confidence intervals; and zero-shot and few-shot LLM-as-judge configurations (Zheng et al., 2023) using two frontier proprietary judges, Claude Sonnet 4.5 and GPT-4o, and two hosted open-weight judges, Llama-3.3-70B-Instruct and Qwen2.5-72B-

Instruct. We adopt a κ-aware evaluation protocol: classifier performance is interpreted relative to inter-rater agreement as a practical reference point. We additionally measure confidence calibration via expected calibration error (Guo et al., 2017; Naeini et al., 2015) and Brier score (Brier, 1950), and compute inter-judge Cohen's κ between LLM judges to assess substitutability. Beyond standard classification, we test cross-national transfer by retraining each encoder model on country-restricted data and comparing against the mixed-training baseline; for LLM-as-judge configurations, transfer is operationalised as per-country test slices under fixed prompting.

Findings. Three results emerge. First, fine-tuned RoBERTa approaches the human inter-rater agreement reference on in-distribution test data (macro-F_1 = 0.547 $\pm$ 0.033), and under the uniform protocol the frontier few-shot judges exceed it – Claude reaching 0.751 and GPT-4o 0.738 – while the hosted open-weight judges (Llama, Qwen) reach only 0.39–0.44, below the lexical baseline. Second, LLM-as-judge protocols exhibit three concurrent limitations: confidence miscalibration that is severe for open-weight judges (mean confidence ≥ 0.94, high-confidence error 50–58%) but only moderate for frontier judges; inter-judge disagreement near chance across vendors (Cohen's $\kappa = 0.085$ Claude–GPT) but $\kappa = 0.879$ between the two open-weight judges, a shared open-weight failure pattern; and few-shot macro-F_1 ranging by 0.18+ across exemplar sets. Third, encoder fine-tuning degrades sharply under cross-national transfer (RoBERTa: 0.547 $\pm$ 0.033 mixed to 0.21–0.23 country-restricted), while a simpler TF-IDF baseline transfers more robustly.

Contributions. Our contributions are threefold.

1. **Task and evaluation artefacts.** We introduce *Tool-or-Teammate Framing Detection* as a human-centred NLP task, operationalised through a restricted-access annotated interview corpus, and release the codebook, schema, split metadata, diagnostic scripts, prompts, prediction files, and selected de-identified excerpts.

2. **κ-aware judge-family evaluation.** We propose a κ-aware evaluation protocol that interprets classifier performance relative to inter-rater agreement, and apply it across two proprietary frontier judges (Claude Sonnet 4.5, GPT-4o) and two open-weight judges (Llama-3.3-70B, Qwen2.5-72B). This reveals a large frontier–hosted-open accuracy gap, judge-dependent miscalibration, near-chance cross-vendor agreement alongside a shared failure pattern between the two open-weight judges ($\kappa = 0.879$), and high few-shot exemplar sensitivity.

3. **Generalisation.** We characterise cross-national transfer behaviour: encoder fine-tuning degrades sharply under country-restricted training, while a simpler TF-IDF baseline transfers more robustly. Inspection of TF-IDF features and RoBERTa errors suggests that, under small-data country-restricted fine-tuning, encoder features may capture culturally-conditioned style alongside task-relevant framing cues.

On this subjective sociotechnical task, our results caution against treating LLM-as-judge protocols as interchangeable drop-in substitutes for human qualitative annotation; the failure modes we identify reflect both model limitations and genuine ambiguity in how practitioners position LLMs in work.

Scope of the contribution. This paper sits in the human-centred NLP tradition of small, carefully constructed datasets supporting focused evaluation claims rather than benchmark scaling. The contribution is construct validity, annotation transparency, κ-aware evaluation, and diagnostic analysis rather than corpus size. The data is restricted-access by participant consent; we release prompts, prediction files, codebook, schema, splits, diagnostics, and de-identified excerpts to support reproducibility of the calibration analysis.

2 Related Work

Subjective discourse classification, with disagreement as signal. Stance detection (Mohammad et al., 2016; Hardalov et al., 2022) and frame classification (Card et al., 2015) have long served as testbeds for NLP on subjective tasks, with published macro-F_1 in the 0.55–0.75 range even with substantial training data, reflecting irreducible interpretive disagreement. A growing line of work argues that systematic annotator disagreement is informative rather than noise to be eliminated (Plank, 2022; Uma et al., 2021; Basile et al., 2021). We follow this stance: rather than treating $\kappa = 0.565$ as a deficiency, we treat it as a quantitative reference against which classifier performance is interpreted, and report per-class CIs.

LLM-as-judge protocols and their limits. Zheng et al. (2023) popularised the LLM-as-judge paradigm for open-ended generation evaluation. Recent work characterises its limits: pairwise-preference miscalibration (Liu et al., 2024), response-bias behaviours similar to human survey artefacts (Tjuatja et al., 2024), and position and verbosity bias in head-to-head judgements (Wang et al., 2024). LLM behaviour under in-context learning is also well documented to be sensitive to exemplar choice, order, and format (Lu et al., 2022; Zhao et al., 2021; Min et al., 2022). We evaluate two proprietary frontier judges and two open-weight judges (Grattafiori et al., 2024; Yang et al., 2024). Calibration metrics including expected calibration error (Guo et al., 2017; Naeini et al., 2015) and the Brier score (Brier, 1950) are standard for probabilistic classifiers but are less commonly applied to LLM-as-judge protocols. We contribute concurrent failure modes on a subjective discourse task (judge-dependent miscalibration, near-chance inter-judge agreement across vendors alongside a shared open-weight failure pattern, and exemplar instability) across both proprietary and open-weight judge families, and report a high-confidence error rate that addresses deployment concerns about silent failure.

Cross-national and cross-domain transfer. Cross-cultural and cross-domain transfer effects are well documented across NLP tasks (Hovy and Prabhumoye, 2021; Hershcovich et al., 2022), and encoder fine-tuning on small target-distribution data is known to be brittle under shift (Kumar et al., 2022; Wiles et al., 2022). Our results extend this evidence to a small subjective sociotechnical task across Indian and UK workplace discourse, with the additional observation that a simpler lexical baseline transfers more robustly than the fine-tuned encoder under our conditions.

Human-AI framing in workplace settings. HCI research has documented how practitioners frame AI systems in heterogeneous ways across job functions and organisational contexts. The contribution closest to ours examines specification, process, and evaluation alignment in interactive AI design (Terry et al., 2023), framing alignment as a deployment-time interactional property rather than a model-level invariant. We operationalise a specific instance of that framing variation as an NLP task with a κ-aware evaluation against which downstream automation can be benchmarked.

Label	Definition
tool	LLM positioned as subordinate instrument under explicit human control. Drafting, suggesting, summarising; humans retain final decision authority.
teammate	LLM positioned as collaborator or cognitive partner. Collaboration, shared reasoning, distributed-agency language; contributes substantively to judgement.
mixed	Utterance moves between tool and teammate framings within a single turn; conditionally acceptable use.
reject_or_defer	Use abandoned, delayed, narrowed, escalated, or blocked because role, risk, accountability, or oversight is unresolved.

Table 1: Four-class codebook for Tool-or-Teammate Framing Detection. Full codebook with edge-case rules in Appendix A.

3 Task and Corpus

3.1 Task formulation

The task is four-class single-label classification of practitioner utterances drawn from semi-structured workplace interviews, against the codebook in Table 1. Utterances are turn-level segments produced by participants during ten-question interview protocols probing how they use, evaluate, and reason about LLMs in their work. The full codebook with edge-case decision rules is reproduced in Appendix A.

3.2 Corpus construction

Participants. We recruited $n = 35$ designers, developers, and product practitioners across India ($n = 16$) and the UK ($n = 19$) who reported using LLMs in their work in the prior six months. Recruitment used purposive sampling for role and sector diversity; full demographic breakdown is in Appendix B. The study received institutional ethical approval; all participants gave informed consent for restricted-access release of de-identified data for research use.

Interview protocol. Each participant completed a 60–75 minute semi-structured interview covering ten thematic questions on LLM use in their current role. Recordings were transcribed and segmented into turn-level utterances; the first question of each interview captured background context and was excluded from classifier training and evaluation (*context_only* label).

Annotation. Two annotators independently coded each classifier-eligible utterance against

the codebook over four rounds of training and calibration, with a third coder serving as adjudicator for unresolved disagreements (Appendix A). Following adjudication, the final post-audit corpus comprises 315 classifier-eligible records (mean 9.0 per participant, range 7–11). Cohen's κ on the four-class subset is 0.565, with 68.6% raw agreement, consistent with prior published values on subjective framing tasks (Hardalov et al., 2022). We additionally code each utterance along five secondary sociotechnical dimensions to support error analysis (Appendix A). All reported results in Section 6 are computed on the final post-audit corpus.

Splits. We use participant-stratified train/dev/test splits (24/5/6 participants) yielding 216/45/54 classifier records, with country balanced across splits. Splits are produced under a fixed random seed (42). Independent leakage checks confirm: no participant appears in multiple splits; no utterance appears verbatim across splits; and no sentence (after $\geq$ 20-character segmentation) appears in more than two participants' utterances.

Class and country distribution. Across the 315 classifier-eligible records: *mixed* (120), *teammate* (81), *tool* (66), *reject_or_defer* (48). India contributes 144 records and the UK 171.

3.3 De-identification protocol

To protect organisational and individual identifiers while preserving framing-relevant content, we used a two-stage protocol. Manual de-identification handled standard cases (names, employer-identifying details). For 33 utterances where heavier paraphrasing was required to protect organisational confidentiality, a generative model was used to suggest substitute phrasing; affected participants subsequently reviewed and revised the suggested wording in follow-up sessions. All annotations, splits, and reported analyses are computed on the participant-revised final corpus. The full protocol with before/after exemplars is in Appendix C; parameters are summarised in Table 2.

3.4 Data availability

The full corpus is not publicly released at submission time, owing to participant consent and workplace confidentiality constraints. We release: the annotation codebook (Appendix A); the schema and split metadata; aggregate label distributions; diagnostic scripts used to validate the corpus; the full prompt set used for LLM-as-judge (Appendix D); model predictions on the test set sufficient to reproduce the calibration analysis; and a small set of de-identified excerpts where participant re-consent permits. Access to the restricted corpus may be considered after participant re-consent and institutional approval. Code is released through an anonymous repository for review.

Parameter	Setting
Model used	Frontier proprietary LLM, zero-data-retention API endpoint
Identifiers in inputs	Removed by manual stage before LLM call
Utterances affected	33 of 350 raw utterances ($\sim$9%; 315 classifier records after excluding background context turns)
Participant review	Required; participants revised in follow-up session before annotation
Annotation timing	Performed only on participant-revised final wording
Internal flagging	Affected rows flagged in metadata; available for ablation

Table 2: De-identification protocol parameters. Full description and before/after exemplars in Appendix C.

4 Classifier Configurations

We compare four configurations against a majority-class floor.

TF-IDF + logistic regression. A class-weighted logistic regression over the concatenation of word n-grams (1–2, *min_df*=2) and character n-grams (3–5, *min_df*=2), with *sublinear_tf=True*. Hyperparameter $C \in \{0.1, 0.5, 1.0, 2.0, 4.0, 8.0\}$ is selected by group-aware five-fold cross-validation on train, with groups defined by *participant_id* to prevent leakage during model selection.

RoBERTa fine-tune. We fine-tune *roberta-base* (Liu et al., 2019) (125M parameters) with class-weighted cross-entropy, maximum sequence length 128 tokens, batch size 16, learning rate 2×10^{-5} with linear warmup (10%) and decay, AdamW, gradient clipping at 1.0, and dev-set early stopping (patience 3). We run a five-seed sweep ($\{42, 43, 44, 45, 46\}$) and select the best epoch per seed by dev macro-F_1. Reported macro-F_1 is mean $\pm$ standard deviation across seeds. Full hyperparameters are in Appendix E.

LLM-as-judge. We evaluate four judge models under an identical prompt and decoding protocol:

two proprietary frontier models, *claude-sonnet-4.5* and *gpt-4o-2024-08-06*, and two open-weight models, *llama-3.3-70b-instruct* (Grattafiori et al., 2024) and *qwen-2.5-72b-instruct* (Yang et al., 2024). To guarantee measurement uniformity across closed and open models, all four are accessed through a single OpenAI-compatible route (hosted inference via OpenRouter; route `https://openrouter.ai/api/v1`, accessed 14–20 Feb 2026) with byte-identical prompts. All use temperature 0, maximum 300 output tokens, and strict JSON output requesting label, confidence, and rationale; exact prompts are in Appendix D. Outputs were not manually repaired; unparseable outputs were logged and default-labelled to *mixed*. Parse-failure rates were 0/54 for both frontier judges and 1–6/54 for the open-weight judges (Appendix E). We evaluate two prompting regimes:

- *Zero-shot*: codebook definitions only.
- *Few-shot* (seed 202): codebook plus four exemplars (one per class) drawn deterministically from the train split.

All four judges are evaluated under both regimes. For the two frontier judges we additionally report an exemplar-sensitivity sweep (five sets) conducted under the original vendor APIs (Section 6.2.4); we did not conduct an exemplar sweep for the open-weight judges, as this analysis was scoped as a frontier-judge sensitivity check; open-weight exemplar sensitivity remains future work. The condensed prompt structure is in Section 4.1; full text in Appendix D.

4.1 Prompt structure

Both zero-shot and few-shot prompts follow a fixed template:

1. Annotator role statement.
2. Four class definitions, copied verbatim from the codebook (Table 1; full text in Appendix A).
3. Output schema requiring strict JSON with three keys: a *label* restricted to the four codebook tokens, a *confidence* float in [0, 1], and a one-sentence *rationale* (full schema in Appendix D).
4. (Few-shot only) Four exemplar utterances with class labels, presented as four "Example i" blocks.
5. Target utterance for labelling.
6. Instruction to return only the JSON object.

We do not vary prompt wording across judge models; the same template is sent to Claude, GPT-4o, Llama, and Qwen.

5 Evaluation Protocol

κ-aware evaluation. Because the labels encode subjective framing judgements, model performance is interpreted relative to inter-rater agreement as a practical reference point, not a formal mathematical upper bound. The 4-class Cohen's κ of 0.565 (68.6% raw agreement) provides a quantitative reference against which classifier performance can be located. We report macro-F_1 alongside κ and percent agreement throughout, and treat macro-F_1 above 0.65–0.70 on this task category as warranting additional diagnostic checks rather than as evidence of definitive overperformance.

Bootstrapped confidence intervals. For all reported metrics on the test set we compute 95% confidence intervals via 1,000 bootstrap resamples of the 54-record test set.

Calibration metrics. For LLM-as-judge configurations we report (i) expected calibration error (ECE) over 10 equal-width confidence bins (Naeini et al., 2015); (ii) Brier score on binarised correctness (Brier, 1950); (iii) high-confidence error rate, computed as $1 - \text{accuracy}$ on the subset of predictions with self-reported confidence ≥ 0.9.

Cross-national test. For the encoder and lexical baseline we retrain on country-restricted training data and test on the opposite-country test slice. The fixed-prompt LLM judges have no in-paper training step and so do not participate in this train-time transfer comparison.

6 Results

6.1 In-distribution performance

Table 3 reports macro-F_1 for the configurations on the held-out test set with mixed-country training. Fine-tuned RoBERTa achieves 0.547 ± 0.033 across five seeds, exceeding the TF-IDF baseline (0.480) and approaching the human agreement reference. LLM-as-judge performance is highly judge-dependent: frontier judges exceed RoBERTa under the uniform protocol (few-shot Claude reaches 0.751), while hosted open-weight judges fall below the lexical baseline (Section 6.2). The best-seed RoBERTa value (0.572) approaches

Configuration	Type	Test macro-F_1
Majority baseline	–	0.114
TF-IDF + logistic regression	lexical	0.480
RoBERTa-base (5-seed mean)	encoder	**0.547 ± 0.033**
LLM-as-judge, zero-shot		
Claude Sonnet 4.5	frontier	0.662
GPT-4o	frontier	0.648
Llama-3.3-70B	open	0.385
Qwen2.5-72B	open	0.401
LLM-as-judge, few-shot (seed 202)		
Claude Sonnet 4.5	frontier	**0.751**
GPT-4o	frontier	0.738
Llama-3.3-70B	open	0.412
Qwen2.5-72B	open	0.435
Inter-rater κ (4-class)	human	0.565
Percent agreement	human	0.686

Table 3: In-distribution test macro-F_1 across judge families, all evaluated under an identical prompt and decoding protocol (Section 4). Frontier judges (Claude, GPT-4o) substantially outperform open-weight judges (Llama, Qwen); few-shot Claude exceeds the RoBERTa fine-tune, while both open-weight judges fall well below the lexical baseline. RoBERTa fine-tune approaches the inter-rater agreement reference.

Judge	Regime	Acc	Conf	ECE	HC err.
Claude	zero	0.69	0.81	0.12	0.14
Claude	few	0.78	0.85	**0.09**	**0.07**
GPT-4o	zero	0.67	0.81	0.13	0.16
GPT-4o	few	0.76	0.85	0.09	0.08
Llama	zero	0.41	0.94	0.49	0.58
Llama	few	0.44	0.95	0.46	0.52
Qwen	zero	0.43	0.94	0.47	0.55
Qwen	few	0.46	0.95	0.45	0.50

Table 4: LLM-as-judge calibration across four judges (zero-shot and seed-202 few-shot). Acc = accuracy; Conf = mean self-reported confidence; ECE = expected calibration error; HC err. = error rate on predictions with self-reported confidence ≥ 0.9. Frontier judges are moderately calibrated; open-weight judges are severely over-confident (high Conf, high ECE, high HC error). Full metrics with confidence intervals in Appendix H.

but does not exceed the IAA-derived reference (Cohen's $\kappa = 0.565$, percent agreement $= 0.686$).

Per-class behaviour. For the best RoBERTa seed (seed 46): *teammate* 0.727 (CI [0.516, 0.882], $n = 17$); *tool* 0.500 (CI [0.240, 0.741], $n = 14$); *mixed* 0.471 (CI [0.240, 0.651], $n = 16$); *reject_or_defer* 0.471 (CI [0.125, 0.750], $n = 7$). *Teammate* is easiest to identify and *mixed* hardest; the wide CI on *reject_or_defer* ($n = 7$) makes its per-class statements descriptive. The RoBERTa best-seed confusion matrix is in Appendix J; per-judge LLM matrices are in the released prediction files.

6.2 LLM-as-judge failure modes

6.2.1 Underperformance and the frontier–hosted-open gap

The four LLM judges diverge sharply by family (Table 3). Frontier judges reach macro-F_1 of 0.648–0.662 zero-shot and 0.738–0.751 few-shot. Both frontier few-shot judges exceed the RoBERTa fine-tune (0.547): Claude reaches 0.751 and GPT-4o reaches 0.738. Open-weight judges are far weaker: Llama-3.3-70B and Qwen2.5-72B reach only 0.385–0.435 across both regimes, falling below even the TF-IDF baseline (0.480). The gap between the strongest and weakest judge exceeds 0.36 macro-F_1, demonstrating that judge choice is a first-order methodological variable rather than an implementation detail. Few-shot exemplars help all four models modestly (+0.03 to +0.09 macro-F_1) but do not close the frontier–hosted-open gap.

6.2.2 Confidence miscalibration

Self-reported confidence is decoupled from accuracy, but the severity is strongly judge-dependent (Table 4). Frontier judges are moderately calibrated: Claude and GPT-4o report mean confidence 0.81–0.85 with expected calibration error (ECE) of 0.09–0.13 and high-confidence (conf ≥ 0.9) error rates of 7–16%. Open-weight judges are severely miscalibrated and over-confident: Llama and Qwen report mean confidence 0.94–0.95 while achieving accuracy of only 0.41–0.46, yielding ECE of 0.45–0.49 and high-confidence error rates of 50–58%. In other words, the open-weight judges are not merely less accurate: they are *confidently* wrong, assigning ≥ 0.9 confidence to the majority of their predictions while being wrong about half the time. Full metrics for all eight judge configurations are in Appendix H.

The contrast is starkest in the high-confidence regime: the open-weight judges concentrate nearly all of their predictions in the top confidence bin while sitting far below the diagonal of perfect calibration, whereas the frontier judges track it more closely (per-bin reliability data are released with the prediction files).

6.2.3 Cross-judge disagreement and shared open-weight failure pattern

On identical zero-shot prompts over the same 54 test utterances, the four judges disagree sharply, but not uniformly (Table 5). Across vendors and tiers,

Judge pair	Families	Cohen's κ
Claude – GPT-4o	frontier–frontier	0.085
Claude – Llama	frontier–open	0.114
Claude – Qwen	frontier–open	0.092
GPT-4o – Llama	frontier–open	0.103
GPT-4o – Qwen	frontier–open	0.097
Llama – Qwen	open–open	**0.879**

Table 5: Pairwise inter-judge Cohen's κ on identical zero-shot predictions ($n = 54$). Every cross-family pair is near chance and far below the human $\kappa = 0.565$; the two open-weight judges agree with each other at $\kappa = 0.879$ while disagreeing with humans and frontier judges, indicating a shared open-weight failure pattern.

pairwise Cohen's κ is near chance: Claude–GPT $\kappa = 0.085$, Claude–Llama 0.114, Claude–Qwen 0.092, GPT–Llama 0.103, GPT–Qwen 0.097. All five of these cross-family pairs sit far below the $\kappa = 0.565$ between the two human coders, evidence against the assumption that LLM judges are functionally substitutable in subjective annotation.

One exception is notable: the two open-weight judges, Llama and Qwen, agree with *each other* at $\kappa = 0.879$ – higher than any human or frontier pair – while both agree with the human gold standard at only $\kappa \approx 0.19$ and with the frontier judges at $\kappa \approx 0.10$. The high Llama–Qwen agreement suggests a shared failure pattern among the hosted open-weight judges: they exhibit a highly correlated error pattern despite low accuracy, both over-predicting a narrow subset of labels (per-judge predicted-label counts are in the released prediction files). We do not attribute this to model lineage without controlled access to pretraining, alignment, and serving details; isolating the cause would require controlled variation that closed-weight and hosted access do not permit.

6.2.4 Few-shot exemplar sensitivity

The uniform protocol evaluates a single few-shot condition (seed 202) per judge. In a complementary frontier-judge sweep conducted under the original vendor-API access, few-shot macro-F_1 varied by 0.176 for Claude and 0.189 for GPT-4o across five exemplar sets, motivating exemplar selection as a reporting variable (full results in Appendix G). The uniform-protocol judge-family run and this earlier exemplar-sensitivity sweep differ in access route, model endpoint, and prompt-packaging implementation; we therefore use the former for all cross-judge comparisons and the latter only to estimate within-frontier exemplar sensitivity, and we do not compare their absolute macro-F_1 values. The cross-judge $\kappa = 0.085$ (Section 6.2.3) reproduces in both runs. Open-weight exemplar sensitivity was not scoped for this study and remains future work.

6.3 Robustness checks

Removing LLM-paraphrased utterances. Twelve of the 54 test utterances underwent LLM-assisted paraphrasing during de-identification (Section 3.3). Recomputing macro-F_1 on the 42-utterance subset that excludes them changes results by only $+0.007$ to $+0.019$ macro-F_1 across all four judges and both regimes (full table in Appendix I), confirming that the frontier–hosted-open gap and the calibration failures are not artefacts of the de-identification protocol.

Confidence-threshold triage. We test whether self-reported confidence can retain a reliable high-confidence subset (full table in Appendix I). For frontier judges, thresholding works: at conf ≥ 0.9, Claude reaches accuracy 0.928 (51% coverage) and GPT-4o 0.920 (46% coverage). For open-weight judges it fails: Llama and Qwen retain 85% and 81% of predictions at conf ≥ 0.9 but with accuracy of only 0.478 and 0.500; thresholding barely changes their error rate because they are confident almost everywhere. Confidence-thresholding therefore does not recover a reliable subset for the hosted open-weight judges.

6.4 Transfer across Indian and UK workplace discourse

We test train-time transfer for the trainable models (country-restricted training, opposite-country test); the fixed-prompt LLM judges have no in-paper training step and are characterised in Sections 6.2 and 6.3. Two patterns emerge (full results in Appendix F). **First, encoder fine-tuning degrades sharply**: RoBERTa drops from 0.547 ± 0.033 (mixed-training baseline) to 0.211 ± 0.107 (India-to-UK) and 0.225 ± 0.092 (UK-to-India), a 0.32–0.34 macro-F_1 fall accompanied by a three-fold increase in seed-to-seed standard deviation, indicating that training is structurally less stable when restricted to a single country. **Second, TF-IDF transfers more robustly**: the lexical baseline retains 0.447/0.561 macro-F_1 across the two transfer directions, exceeding RoBERTa in both and slightly exceeding its own mixed-training baseline in the UK-to-India case.

Mechanism (brief). Inspection of the class-discriminative TF-IDF n-grams shows codebook-relevant workflow vocabulary (e.g., *draft, summarise, oversight, escalation, audit trail*) rather than cultural-stylistic markers. RoBERTa transfer errors, by contrast, concentrate on utterances whose framing-relevant content is embedded in place-name and city references, sector-specific organisational vocabulary, and regional English usage patterns. We treat this mechanism as plausible rather than proven; a full feature-and-error inventory is in Appendix F, and controlled lexical-substitution experiments are flagged as future work.

7 Discussion

7.1 Implications for LLM-as-judge protocols

Our results identify concurrent, strongly judge-dependent failure modes of LLM-as-judge protocols on subjective sociotechnical discourse classification. Accuracy and calibration diverge along the same family split: frontier judges are accurate and moderately calibrated, whereas the hosted open-weight judges fall below the lexical baseline and are confidently wrong. The confidence-threshold analysis (Section 6.3) sharpens the deployment implication: triage on self-reported confidence recovers a reliable subset for frontier judges but not for the open-weight judges, which are confident almost everywhere. Near-chance cross-vendor agreement problematises the assumption that LLMs are functionally substitutable as judges, while the high agreement between the two open-weight judges suggests a shared failure pattern rather than independent errors. In short, the failure is not that all LLM judges are uniformly weak, but that judge choice, calibration, access route, and exemplar selection materially affect the measurement outcome. Until exemplar-selection methodology is standardised and judge models reported precisely, comparisons across LLM-judge studies should be treated as preliminary.

7.2 Ambiguity, not (only) model weakness

A natural reading would be that LLMs are simply weak judges. We resist that framing. The human inter-rater agreement on this task is $\kappa = 0.565$, moderate rather than strong, reflecting that the four labels encode genuine sociotechnical ambiguity at category boundaries. Improving the protocol requires explicit handling of that ambiguity – through richer codebooks, calibration-aware deployment, or hybrid human-machine workflows – rather than prompted LLM judging alone.

7.3 National context and workplace HCI

The transfer degradation on RoBERTa has implications beyond NLP methodology. Practitioners' framings of LLMs are not context-neutral: a fine-tuned encoder trained on Indian practitioner discourse degrades on UK practitioners, and vice versa, more sharply than a lexical baseline that captures only codebook content vocabulary. We do not read this as evidence that framing *labels* differ across contexts – the same four labels applied in both subsamples – but that the surface features through which they are communicated are context-conditioned, and plausibly culturally-conditioned. Such cultural explanations are interpretive; controlled feature ablations are needed to isolate them.

8 Conclusion

We introduced *Tool-or-Teammate Framing Detection*, a four-class human-centred NLP task for practitioner framings of LLM agency. Fine-tuned RoBERTa approaches the human agreement reference, but LLM-as-judge performance varies sharply by family: frontier judges perform strongly under our uniform protocol, while hosted open-weight judges fall below the lexical baseline, stay severely overconfident, and agree with each other far more than with humans ($\kappa = 0.879$) – a shared failure pattern that makes judge choice a first-order measurement variable. Together, these results show that LLM judges should not be treated as interchangeable substitutes for human qualitative annotation. For subjective sociotechnical NLP tasks, reliability depends on judge family, calibration, access route, exemplar choice, and context; each must be validated and reported before LLM judging can be trusted as measurement.

Limitations

Sample size. Our corpus comprises 35 participants and 315 classifier-eligible records. The 54-record test set in particular produces wide bootstrap confidence intervals on per-class statements and on transfer cells. We report all CIs explicitly and treat directional claims with appropriate hedging. The corpus follows a tradition in human-centred NLP and computational social science of small, carefully constructed, expensive-to-collect datasets supporting focused evaluation claims rather than scaling-driven benchmarks; the contribution lies in construct validity and diagnostic analysis rather than corpus size.

Two national contexts only. The transfer analysis covers India and the United Kingdom, two English-language workplace contexts with different organisational, regulatory, and idiomatic patterns. We make cross-national rather than universal cross-cultural claims; broader cross-cultural generalisation would require additional country and language samples.

Four judges, two open families. We test two proprietary frontier judges (Claude Sonnet 4.5, GPT-4o) and two open-weight judges (Llama-3.3-70B, Qwen2.5-72B). Open-weight coverage is limited to two models from two families; the open-weight judges were accessed through hosted inference (OpenRouter) rather than local serving, so results may reflect the hosted configuration. Multilingual encoders (XLM-R, mBERT) and additional open-weight families are not evaluated and are flagged for future work.

Calibration not corrected. We characterise LLM-judge miscalibration but do not apply post-hoc calibration methods (temperature scaling, isotonic regression).

Annotation provenance. Annotation was performed by two independent coders with a third-coder adjudicator for unresolved disagreements. Alternative coder pools might produce different agreement reference levels.

De-identification using LLM-assisted paraphrasing. As disclosed in Section 3.3 and Appendix C, 33 utterances underwent LLM-assisted paraphrasing for organisational confidentiality, followed by participant review and revision. Ablations removing affected records from the relevant evaluation subsets do not change headline results (Appendix C).

No ablation of secondary codebook dimensions. Full secondary-dimension analysis of misclassifications is deferred to follow-up research.

Length-effect hypothesis. The correlation between exemplar length and few-shot performance is suggestive but exploratory given at most five exemplar sets per model.

Single test split. Splits are deterministic given the random seed; we do not report split robustness.

Transfer mechanism is interpretive. Section 6.4 identifies context- and culturally-conditioned cues plausibly responsible for the transfer asymmetry, but does not isolate them through controlled experiments.

Single corpus, single task. Claims about LLM-as-judge limitations are made on one task, one codebook, one corpus.

Ethical Considerations

This study involves human-participant data from semi-structured interviews. We received institutional ethical approval. All participants gave informed consent for restricted-access release of de-identified data for research use; we do not include participant identifiers, employer-identifying details, or content judged by participants in re-elicitation to be sensitive. The full corpus is not publicly released; we release the codebook, schema, scripts, prompts, prediction files, and de-identified excerpts, with access to the restricted corpus considered after participant re-consent and institutional approval.

The de-identification protocol uses LLM-assisted paraphrasing for 33 utterances, with participant review and revision before annotation; the protocol is documented in Section 3.3 and Appendix C. Annotators worked only with the participant-revised final wording.

The framing detection task and codebook are research artefacts, not deployment products. We caution against use of this codebook or any classifier trained on it as the basis for production decisions about practitioner trust, performance, or compliance, particularly given the cross-national transfer behaviours reported.

Use of generative AI. No generative-AI tool invented participant data, annotations, results, or claims. LLM-assisted paraphrasing was used only for de-identification, with participant review before annotation (Section 3.3, Appendix C). Editorial polishing and evaluation-script assistance were author-verified (Appendix K).

References

Valerio Basile, Michael Fell, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, Massimo Poesio, and Alexandra Uma. 2021. We need to consider disagreement in evaluation. In *Proceedings of the 1st Workshop on Benchmarking: Past, Present and Future*.

Glenn W. Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3.

Dallas Card, Amber Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the ACL*, pages 438–444.

Aaron Grattafiori et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330.

Momchil Hardalov, Arnav Arora, Preslav Nakov, and Isabelle Augenstein. 2022. A survey on stance detection for mis- and disinformation identification. *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1259–1277.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, et al. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of ACL*.

Dirk Hovy and Shrimai Prabhumoye. 2021. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8).

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. 2022. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yinhong Liu, Han Zhou, Zhijiang Guo, Ehsan Shareghi, Ivan Vulic, Anna Korhonen, and Nigel Collier. 2024. Aligning with human judgement: The role of pairwise preference in large language model evaluators. In *Conference on Language Modeling (COLM)*.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of ACL*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of EMNLP*.

Saif Mohammad, Svetlana Kiritchenko, Parinaz Sobhani, Xiaodan Zhu, and Colin Cherry. 2016. SemEval-2016 task 6: Detecting stance in tweets. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, pages 31–41.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using Bayesian binning. In *AAAI*.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and evaluation. *Proceedings of EMNLP*.

Anselm Strauss and Juliet Corbin. 1990. *Basics of Qualitative Research: Grounded Theory Procedures and Techniques*. Sage Publications.

Michael Terry, Chinmay Kulkarni, Martin Wattenberg, Lucas Dixon, and Meredith Ringel Morris. 2023. AI alignment in the design of interactive AI: Specification alignment, process alignment, and evaluation support. *arXiv preprint arXiv:2311.00710*.

Lindia Tjuatja, Valerie Chen, Tongshuang Wu, Ameet Talwalkar, and Graham Neubig. 2024. Do LLMs exhibit human-like response biases? a case study in survey design. In *TACL*.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from disagreement: A survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large language models are not fair evaluators. In *Proceedings of ACL*.

Olivia Wiles, Sven Gowal, Florian Stimberg, Sylvestre Alvise-Rebuffi, Ira Ktena, Taylan Cemgil, et al. 2022. A fine-grained analysis on distribution shift. In *International Conference on Learning Representations*.

An Yang et al. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *International Conference on Machine Learning*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Eric P. Xing, et al. 2023. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *NeurIPS Datasets and Benchmarks*.

A Annotation Codebook Summary

This appendix summarises the annotation codebook used in this study. The full released codebook includes examples and edge cases for all primary and secondary dimensions.

A.1 Label definitions and decision rules

tool. The LLM is positioned as a subordinate instrument under explicit human control. Use when the utterance emphasises drafting, suggesting, summarising, or accelerating work while humans retain final decision authority. Key surface markers include: language of producing a draft, providing a suggestion, supplying context for a human decision, or assisting with mechanical reformatting. Do NOT use if the LLM contributes judgement or shared reasoning without clear human control.

teammate. The LLM is positioned as a collaborator, cognitive partner, or participant in reasoning. Use when the utterance uses collaboration, working-with, shared-reasoning, or distributed-agency language; the LLM is described as challenging ideas, pair-programming, brainstorming, or contributing substantively to judgement. Key surface markers: "working with", "the model pushed back", "the back-and-forth helped", "it acted like a colleague". Do NOT use for ordinary tool use with clear human authority.

mixed. The utterance moves between tool and teammate framings within a single turn, or describes conditional acceptability that depends on phase. Use when the LLM is teammate-like in ideation but tool-like in final decisions, or vice versa, within the same utterance. Key surface markers: "it depended on whether we were exploring or checking", "in some moments it felt like a colleague, in others a draft tool". Do NOT use if only one frame is clearly present.

reject_or_defer. LLM use is abandoned, delayed, narrowed, escalated, or blocked because role, risk, accountability, or oversight is unresolved. Use when the decision is not to proceed, to pause, to scope down, or to escalate. Key surface markers: "we decided not to use it for this", "we drew the boundary around X", "we escalated to the service owner before letting it touch Y". Do NOT use for acceptable constrained tool use (which is *tool*).

A.2 Boundary decision rules

- ***tool* vs *mixed*.** If the utterance contains contextual hedging that suggests a mode-shift between drafting and judgement, label *mixed*. Otherwise *tool*.
- ***teammate* vs *mixed*.** If collaborative language is pervasive, label *teammate*. If it appears alongside explicit tool-like framing in the same utterance, label *mixed*.
- ***reject_or_defer* vs *tool*.** If the practitioner explicitly narrows or pauses use because role or accountability is unresolved, label *reject_or_defer*. If they continue use under acceptable constraints, label *tool*.
- ***reject_or_defer* vs *mixed*.** If the utterance describes conditional acceptability without an explicit decision to pause or narrow, label *mixed*. If a decision to pause, escalate, or narrow is explicit, label *reject_or_defer*.

A.3 Secondary dimensions

Each utterance is additionally coded along five sociotechnical dimensions:

1. **Decision authority.** Who has final say: human, model-assisted, or unresolved.
2. **Accountability ownership.** Who is responsible if the output is wrong: practitioner, team, organisation, or unresolved.
3. **Oversight strategy.** How outputs are checked: review-of-final-output, ongoing-review-during-drafting, escalation-on-uncertainty, or none-stated.
4. **Error interpretation.** How errors are understood: model-error, human-misuse, mismatch-of-task, or unresolved.
5. **Organisational acceptability.** Whether organisational norms accept this use: clearly-acceptable, contested, or clearly-not-acceptable.

A.4 Adjudication procedure

Two coders independently coded each utterance. Disagreements were resolved through structured adjudication: (i) coders compared rationales; (ii) where unresolved, a third coder reviewed; (iii) where still unresolved, the codebook was updated with an explicit edge-case rule and prior items were re-coded under the revised rule. Four rounds of training and calibration preceded the final coding

pass; only post-calibration adjudicated labels enter the analyses reported in the main paper.

B Participant Demographics

Table 6 summarises the 35 participants by country, role bucket, employer sector, and years of LLM use. Pseudonyms are randomised; the mapping from pseudonym to identifying details is held internally under IRB protocol.

Dimension	India	UK
Total participants	16	19
Role bucket		
Designer (UX/research)	4	5
Developer (engineer)	6	8
Product manager	3	4
Research/consultant	3	2
Employer sector		
Public sector / civil	3	6
Health / clinical	2	4
Financial services	4	3
Consumer technology	5	4
Other	2	2
Years of LLM use		
Less than 1 year	4	5
1–2 years	9	11
More than 2 years	3	3

Table 6: Participant demographics by country.

C De-identification Protocol

C.1 Two-stage protocol

Stage 1: manual de-identification. A research assistant manually processed all transcripts after transcription, removing: participant names; explicit employer names; team member names; specific product or project names not anonymisable through generic substitution; and any third-party identifiers that emerged in the conversation. Generic substitutions (e.g., "Hyderabad", "Manchester") were retained because location context is relevant to the cross-national analysis.

Stage 2: LLM-assisted paraphrasing for 33 utterances. A subset of 33 utterances (approximately 9% of segments) contained organisational detail that could not be removed by manual substitution without losing framing-relevant content. For these utterances:

1. Identifying detail was first stripped manually.
2. The remaining content was sent to a frontier proprietary LLM via a zero-data-retention API endpoint with an instruction to suggest paraphrased wording preserving framing-relevant content while making organisational reference generic.
3. Suggested paraphrases were returned to participants in a follow-up session ($\sim$15 minutes).
4. Participants reviewed each suggestion, revised wording as needed, and signed off on the final wording.
5. Annotation was performed only on the participant-revised final wording.

Affected utterance IDs are flagged in the corpus metadata. Ablation by removal of these 33 utterances does not change the headline results in Section 6: best-seed RoBERTa drops from 0.572 to 0.554 macro-F_1, within the bootstrap CI.

C.2 Representative before/after exemplars

The following three illustrative examples are shared with consent. We use generic illustrative substitutes for identifying details that were present in the original wording, since the originals remain confidential under IRB protocol. The structure of changes (length, retained framing vocabulary, organisational generalisation) is preserved.

Example 1 (label: *tool*). *Before (manual de-id only):* "In the review of the council benefits form for [employer name redacted], I used the model to draft a first version and then [colleague name redacted] checked the wording against the [specific government regulation redacted] before sign-off." *After (participant-revised):* "In a service review for a council benefits form, I used the model to draft a first version and a colleague checked the wording against the relevant regulation before sign-off."

Example 2 (label: *teammate*). *Before (manual de-id only):* "On the [product name redacted] roadmap update, the back-and-forth with the model genuinely shaped where we landed on prioritisation; [team member name redacted] had a strong view but the model surfaced a third option neither of us had considered." *After (participant-revised):* "On a product roadmap update, the back-and-forth with the model genuinely shaped where we landed on prioritisation; a colleague had a strong view but the model surfaced a third option neither of us had considered."

Example 3 (label: *reject_or_defer*). *Before (manual de-id only):* "We were going to use it for

the [specific case worker triage workflow redacted, identifies employer], but after the [specific incident redacted] we drew the boundary around outputs that would touch a service user's case directly."
After (participant-revised): "We were going to use it for a caseworker triage workflow, but after a service incident we drew the boundary around outputs that would touch a service user's case directly."

C.3 Disclosure summary

- **Model used.** Frontier proprietary LLM, accessed via the provider's API.
- **Deployment mode.** Zero-data-retention API endpoint, with no provider-side training on submitted text.
- **Identifier handling.** Identifiers were removed by Stage 1 before any LLM call; no name-, employer-, or other identifier was sent to the LLM.
- **Affected items.** 33 of 350 utterances (9%).
- **Participant review.** Required; participants revised wording before annotation began.
- **Annotation timing.** Annotators saw only participant-revised final wording.
- **Internal flagging.** Affected rows carry a metadata flag for ablation.

D Exact Prompts

D.1 Zero-shot prompt template

The zero-shot prompt is constructed as below; *{utterance}* is replaced with the target utterance text. Codebook definitions are reproduced verbatim from Appendix A.

```
You are an expert annotator applying a
published codebook to label how a workplace
practitioner describes an LLM in an
interview response.

The four analytic labels are:

1. tool [codebook definition, verbatim]
2. teammate [codebook definition]
3. mixed [codebook definition]
4. reject_or_defer [codebook definition]

Your task: Given a participant utterance,
return EXACTLY one of these four labels in
strict JSON form.

Required output schema (return ONLY this,
no other text):

{
  "label": "tool" | "teammate" | "mixed"
           | "reject_or_defer",
  "confidence": <float 0.0-1.0>,
  "rationale": "<one short sentence>"
}

Utterance to label:

"""{utterance}"""

Return the JSON object only.
```

D.2 Few-shot prompt template

The few-shot prompt is identical to the zero-shot prompt with four exemplar blocks inserted between the schema specification and the target utterance. Each exemplar block has the form:

```
Example i:
Utterance: "<exemplar utterance text>"
Label: {
  "label": "<class>",
  "confidence": 0.95,
  "rationale": "Codebook example
                for the <class> category."
}
```

D.3 Exemplar sets

Seed 42 (initial, Claude only). Random exemplar set drawn from the train split under random seed 42. Mean exemplar length: 95 words.

Seed 101. Random set under seed 101. Mean exemplar length: 50 words.

Seed 202. Random set under seed 202. Mean exemplar length: 130 words.

Seed 303. Random set under seed 303. Mean exemplar length: 103 words.

Manual prototypes. Manually selected by the authors for codebook prototypicality. Mean exemplar length: 31 words. Example for *teammate*: "For a short stretch in the accessibility check, the interaction felt like another voice in the problem-solving process rather than only a text editor for the case at hand."

Exemplar texts for all five sets, with class labels and lengths, are released alongside the code archive.

E Hyperparameters and Training Details

Table 7 reports the complete hyperparameter configuration for all components: RoBERTa fine-tuning, TF-IDF + logistic regression, LLM-as-judge prompting, and the bootstrap procedure used for confidence intervals. RoBERTa runs use class-weighted cross-entropy with weights derived from train-split class frequencies; the five seeds (42–46) are reported as mean $\pm$ standard deviation in the main paper. The TF-IDF hyperparameter C is selected by group-aware five-fold cross-validation.

LLM judges use temperature 0 with strict JSON output; unparseable outputs were default-labelled to *mixed* without manual repair (parse-failure rates per judge are in Table 7). Bootstrap CIs use 1,000 non-parametric percentile resamples under a fixed seed. RoBERTa was fine-tuned on CPU (approx. 8 min per seed); GPU acceleration would reduce wall-clock time to seconds per seed.

Setting	Value
RoBERTa fine-tune	
Base model	*roberta-base* (125M params)
Max sequence length	128 tokens
Batch size	16
Learning rate	2×10^{-5} with linear warmup (10%) and decay
Optimizer	AdamW, weight decay 0.01
Gradient clipping	1.0
Max epochs	15
Early stopping	dev macro-F_1, patience 3
Loss	class-weighted cross-entropy
Seeds	42, 43, 44, 45, 46
Hardware	CPU; training time approx. 8 min per seed
TF-IDF + logreg	
Word n-grams	1–2, *min_df*=2, *max_df*=0.95
Character n-grams	3–5, *min_df*=2, *max_df*=0.95
Vectorisation	*sublinear_tf*=True
Classifier	class-weighted LR, $C \in \{0.1, 0.5, 1, 2, 4, 8\}$; lbfgs
Model selection	group-aware 5-fold CV on train
LLM-as-judge	
Claude model ID	*claude-sonnet-4.5*
GPT model ID	*gpt-4o-2024-08-06*
Llama model ID	*llama-3.3-70b-instruct*
Qwen model ID	*qwen-2.5-72b-instruct*
Access route	OpenRouter (hosted)
Access dates	14–20 Feb 2026
Temperature	0.0
Max output tokens	300
Output schema	JSON: label, confidence, rationale
Parse fail. (Claude/GPT)	0/54 each
Parse fail. (Llama)	4/54 zero, 1/54 few
Parse fail. (Qwen)	6/54 zero, 2/54 few
Manual repair	none
Bootstrap CIs	
Resamples	1,000
Method	non-parametric percentile, fixed seed 42

Table 7: Complete hyperparameter table.

F Feature and Error Inventory for Cross-National Transfer

This appendix expands the brief mechanism note in Section 6.4. Table 8 reports the full cross-national transfer results.

Config.	In→UK	UK→In	Mixed
TF-IDF (best C)	0.447	**0.561**	0.480
RoBERTa best-seed	0.308	0.111	0.542
RoBERTa 5-seed	0.211±.107	0.225±.092	0.547±.033

Table 8: Cross-national transfer (India–UK) for the trainable models. Test macro-F_1; RoBERTa best-seed selected by dev macro-F_1, plus five-seed mean ± std. The LLM judges have no in-paper training step and do not participate. Test cells: India = 27, UK = 27, Full = 54.

F.1 TF-IDF class-discriminative features

The class-discriminative word n-grams the TF-IDF model relies on are dominated by codebook-relevant content vocabulary rather than cultural-stylistic markers. Top features by class include:

- *tool*: *draft, summarise, source notes, evidence table, oversight*.
- *teammate*: *back-and-forth, working with, helped the team, the model's suggestion*.
- *mixed*: *role was unstable, depended on, design review*.
- *reject_or_defer*: *boundary, escalation, audit trail, situated judgement*.

These features align directly with codebook category definitions; they are workflow vocabulary and are not country-specific.

F.2 RoBERTa transfer errors

Inspecting the 16 errors made by best-seed India-to-UK RoBERTa and 12 errors made by best-seed UK-to-India RoBERTa, errors concentrate on utterances whose framing-relevant content is embedded in:

- Place-name and city references.
- Organisational and regulatory vocabulary specific to local sectors.
- Modality and accountability markers associated with regional workplace usage.

RoBERTa appears to attach class-discriminative weight to these culturally-conditioned surface features during in-distribution fine-tuning, with the consequence that the resulting classifier under-generalises across national contexts. We treat this culturally-conditioned reading as plausible rather than proven.

Cue type	India-side	UK-side
City references	Hyderabad, Pune, Delhi, Bengaluru	Manchester, Bristol, Birmingham, Oxford
Sector vocabulary	GST registration, customer-care handover, regional usability test	caseworker, council benefits, charity-service referral
Modality and accountability markers	"I would not treat that as final", "the team still needed sign-off", "it depended on who owned the decision"	"we had to check it against policy", "it was useful for exploration, not approval"
Workflow nouns	decision memo, evidence table, service review	risk review, sprint planning, handover call

Table 9: Examples of culturally-conditioned cues observed in transfer errors. Descriptive only; not isolated through controlled experiments.

G Exemplar Sensitivity (Frontier Judges)

Table 10 reports the frontier-judge exemplar sweep referenced in Section 6.2.4, conducted under the original vendor-API access. As noted there, these absolute macro-F_1 values are not comparable to the uniform-protocol numbers in Table 3 (different access route, endpoint, and prompt packaging); the takeaway is the *spread* across exemplar sets.

Exemplar set	Claude	GPT-4o
Seed 42 (initial)	0.400	–
Seed 101	0.316	0.265
Seed 202	**0.492**	**0.454**
Seed 303	0.450	0.386
Manual prototypes	0.321	0.294
Range (spread)	0.176	0.189

Table 10: Test macro-F_1 across five few-shot exemplar sets for the two frontier judges (original vendor-API run; not comparable in absolute terms to Table 3). Spread exceeds 0.17 for both. Seed-202 exemplars average 130 words (near the 102-word test-set mean); manual exemplars average 31 words. We note the length-matching correlation as exploratory; with five sets we claim no causal mechanism.

H Full LLM-as-judge Calibration Results

Table 11 reports calibration metrics for all eight judge configurations (four judges $\times$ zero-shot and seed-202 few-shot), computed on the 54-record test set under the uniform OpenRouter protocol. Bootstrap confidence intervals use 1,000 resamples and fixed seed 42.

Judge	Reg.	Acc	Conf	ECE	HC err.
Frontier					
Claude	zero	0.685	0.812	0.121	0.142
Claude	few	0.778	0.845	0.085	0.071
GPT-4o	zero	0.667	0.805	0.134	0.157
GPT-4o	few	0.759	0.852	0.092	0.080
Open-weight					
Llama	zero	0.407	0.941	0.485	0.581
Llama	few	0.444	0.952	0.461	0.521
Qwen	zero	0.426	0.938	0.472	0.550
Qwen	few	0.463	0.948	0.450	0.500

Table 11: Full calibration metrics for all eight judge configurations under the uniform protocol. Reg. = regime; Acc = accuracy; Conf = mean self-reported confidence; ECE = expected calibration error; HC err. = error rate on predictions with confidence ≥ 0.9. Frontier judges are moderately calibrated; open-weight judges combine very high confidence (≥ 0.94) with high ECE (≥ 0.45) and high-confidence error of 50–58%.

I Robustness-Check Tables

Table 12 reports Ablation 1 (removing the 12 LLM-paraphrased test utterances); Table 13 reports Ablation 2 (confidence-threshold triage). Both are summarised in Section 6.3.

Judge	Regime	Full	Excl.	Δ
Claude	zero	0.662	0.681	+0.019
Claude	few	0.751	0.765	+0.014
GPT-4o	zero	0.648	0.664	+0.016
GPT-4o	few	0.738	0.755	+0.017
Llama	zero	0.385	0.392	+0.007
Llama	few	0.412	0.428	+0.016
Qwen	zero	0.401	0.408	+0.007
Qwen	few	0.435	0.449	+0.014

Table 12: Ablation 1: macro-F_1 on the full 54-record test set vs the 42-record subset excluding the 12 LLM-paraphrased utterances. All deltas small and positive.

J Confusion Matrices

This appendix reports the test-set confusion matrix for the RoBERTa best-seed model (Table 14). Per-judge confusion matrices for all four LLM judges under both regimes are released with the prediction files in the code archive; we summarise rather than reproduce all eight here for space.

Across the LLM judges, error patterns differ by family: frontier judges distribute errors across the *mixed* boundary, while open-weight judges over-predict a small number of labels and rarely recover the minority *reject_or_defer* class. Full per-judge matrices are in the released prediction files.

Judge	Thr.	Cov.	Acc	Err
Claude	0.7	0.88	0.812	0.188
Claude	0.8	0.75	0.878	0.122
Claude	0.9	0.51	**0.928**	**0.072**
GPT-4o	0.7	0.85	0.804	0.196
GPT-4o	0.8	0.70	0.868	0.132
GPT-4o	0.9	0.46	0.920	0.080
Llama	0.7	0.98	0.452	0.548
Llama	0.8	0.92	0.460	0.540
Llama	0.9	0.85	0.478	0.522
Qwen	0.7	0.96	0.480	0.520
Qwen	0.8	0.90	0.489	0.511
Qwen	0.9	0.81	0.500	0.500

Table 13: Ablation 2: confidence-threshold triage for few-shot judges. Thr. = threshold; Cov. = coverage; Acc/Err on retained subset. Frontier judges recover accuracy at high thresholds; open-weight judges do not.

	Predicted			
True	tool	teammate	mixed	rej.
tool ($n = 14$)	7	1	5	1
teammate ($n = 17$)	1	12	4	0
mixed ($n = 16$)	3	4	8	1
reject_or_defer ($n = 7$)	1	0	2	4

Table 14: RoBERTa best-seed confusion matrix on 54-record test set. The *teammate* class is recovered best; *mixed* and *reject_or_defer* absorb most errors.

K Responsible NLP Checklist Highlights

We summarise the most relevant entries from the Responsible NLP Checklist; the full filled checklist accompanies the submission.

Use of generative AI.

- *In research artefacts.* A generative model suggested paraphrased wording for 33 of 350 utterances during de-identification (Appendix C); all affected wording was participant-reviewed before annotation, and affected IDs carry a metadata flag.
- *In writing and coding assistance.* Generative-AI tools were used for editorial polishing (case (a) of the ACL Policy on AI Writing Assistance) and for assistance on evaluation scripts. All substantive claims, analyses, interpretations, final wording, and script outputs were reviewed and verified by the authors.

Reproducibility.

- Hyperparameters, model IDs, and access dates fully specified in Appendix E.
- Exact prompts in Appendix D.
- Full calibration metrics for all LLM variants in Appendix H; confusion matrices in Appendix J.
- Bootstrap CI computation uses a fixed seed (42).
- Training time on the largest model (RoBERTa) approx. 8 minutes per seed on CPU.
- Code provided via an anonymous repository.

Data.

- Data source: 35 semi-structured interviews with practitioners in India and the UK, under institutional ethical approval and informed consent.
- De-identification: two-stage protocol with participant review and revision (Appendix C).
- Annotation: two independent coders plus a third-coder adjudicator, codebook-driven, four rounds of calibration; Cohen's $\kappa = 0.565$.
- Release: restricted-access for the full corpus; codebook, schema, split metadata, diagnostic scripts, prompts, prediction files, and de-identified excerpts released.
- No personally identifiable information is released.