# Whose Pragmatics? Cultural Grounding as a Bottleneck for Stereotype Detection in Egyptian Arabic Social Media

**Samar A. Assem**
Department of Phonetics & Linguistics
Alexandria University
Alexandria, Egypt
samar.assem@alexu.edu.eg

## Abstract

Stereotype detection benchmarks assume that stereotyping occurs through what is said — via lexical co-occurrence between demographic terms and stereotypical attributes. We argue that stereotyping is often conveyed by what is meant: through presupposition, implicature, and speech-act framing that leave surface content unchanged while embedding prejudice in the pragmatic layer. We call this phenomenon *pragmatic stereotyping*. Evaluating GPT-4 and Claude 3.5 Sonnet on a stratified sample of 500 Egyptian Arabic social media comments annotated with a seven-tag sentiment/(im)politeness taxonomy, we find that cultural grounding is the critical bottleneck in detecting pragmatic stereotyping in non-English discourse. About 35% of LLM errors result from cultural grounding gaps, leading to a 15-percentage-point F1 difference between explicit tags (0.81) and implicit tags (0.66). These failures are bidirectional: on the author side, LLMs under-detect prejudice encoded through concessive presupposition and backhanded compliments; on the model side, LLMs apply English-based pragmatic assumptions, misinterpreting genuine polite criticism as sarcasm and positive-intended impoliteness as conflictive. Our five-layer Chain-of-Thought diagnostic framework localizes these failures to the culture-dependent inference layers. These results extend stereotype evaluation beyond lexical benchmarks and have direct implications for content moderation pipelines serving Arabic-speaking communities.

## 1 Introduction

Stereotype detection in large language models is overwhelmingly studied as a lexical phenomenon. Benchmarks such as StereoSet, CrowS-Pairs, and BBQ test whether models associate demographic groups with stereotypical attributes through word-level co-occurrence: women with emotional, Arabs with aggressive, elderly with frail. These benchmarks have driven important progress, but they share a structural limitation; they assume that stereotyping is carried by what is said. In natural discourse, stereotyping is often conveyed through what is meant, via presupposition, implicature, and speech-act framing, which leaves the surface content intact while encoding prejudice in the pragmatic layer.

For example, in the Egyptian Arabic comment "والله برافو عليكي أول مرة أشوف ست بتفهم" ("Bravo, it is my first time to see a woman who understands"), no negative lexical item appears, the sentiment is surface-positive, and a content-level stereotype detector returns clean. Yet, the presupposition triggered by *my first time* and *a woman who understands* encodes the speaker's belief that women generally do not understand. Accordingly, the stereotype lives in the pragmatic structure, not the lexicon, and current evaluation paradigms cannot see it.

We call this phenomenon *pragmatic stereotyping*: stereotyping or bias conveyed through pragmatic mechanisms rather than explicit lexical content. We argue that cultural grounding, which is the ability to recover culture-specific pragmatic baselines for interpreting (im)politeness, sincerity, and social intent, is the critical bottleneck preventing LLMs from detecting pragmatic stereotyping in non-English discourse. To test this claim, we evaluate two state-of-the-art LLMs (GPT-4 and Claude 3.5 Sonnet) on a 500-comment stratified sample of Egyptian Arabic social media. This sample is annotated with a seven-tag



*Proceedings of the 1st Workshop on Stereotypes Across Cultures in Language Technologies (StereACuLT 2026)*, pages 69–78

taxonomy that jointly encodes sentiment and (im)politeness.

Egyptian Arabic is a particularly diagnostic test case for two reasons. First, stereotyping in Egyptian online discourse is often conveyed through pragmatic mechanisms such as concessive presupposition, backhanded compliments, and grudging praise, which leave surface sentiment positive. Moreover, the (im)politeness system itself diverges from English norms, with mock-impoliteness conventionally signaling warmth and sincere polite criticism functioning as a productive speech-act category. Second, the language is low-resource for pragmatic NLP, meaning that LLMs' priors are disproportionately shaped by English pragmatic norms, making the cultural grounding gap empirically visible.

Our evaluation reveals two complementary failures rooted in the same gap. On the author side, LLMs under-detect pragmatic stereotyping since comments encoding prejudice through positive-sentiment surface forms are misclassified as playful banter. On the model side, LLMs impose English-derived pragmatic priors on Egyptian Arabic (im)politeness, misreading sincere polite (mitigated) criticism as sarcasm and positive-intended impoliteness (teasing) as hostility. Approximately 35% of all LLM errors in our evaluation stem from cultural grounding gaps. Additionally, the aggregate F1 gap between tags where form and meaning align (0.81) and tags where they diverge (0.66), a 15-percentage-point difference, quantifies the bottleneck at the task level.

Our contributions are: (1) we introduce the concept of *pragmatic stereotyping* and distinguish it from lexical stereotyping, aiming to expand the scope of stereotype assessment beyond content-level benchmarks; (2) we provide the first empirical evaluation of LLMs on pragmatically-encoded stereotyping in Arabic, using a culturally grounded annotation framework validated by both expert annotators ($\kappa = 0.78$) and 200 native speakers; (3) we demonstrate that failures in cultural grounding cause bidirectional safety miscalibrations, both under-detecting actual stereotyping and over-attributing stereotyped readings to culturally normative (im)politeness, with our SICoT diagnostic framework localizing these failures to culture-dependent inference layers.

## 2 Related Work

Three research dimensions intersect on this paper's contribution: stereotype evaluation in LLMs, pragmatic reasoning in LLMs, and sentiment and social meaning within the broader landscape of Arabic NLP.

**Stereotype evaluation.** Early and current bias benchmarks consider stereotyping as lexical co-occurrence between demographic terms and stereotypical attributes. StereoSet (Nadeem et al., 2021) and CrowS-Pairs (Nangia et al., 2020) measure preference between stereotypical and anti-stereotypical sentence completions. Moreover, BBQ (Parrish et al., 2022) extends this to question-answering with ambiguous and disambiguated contexts. These paradigms have established that LLMs encode social biases, but they evaluate stereotyping only where it surfaces lexically. Blodgett et al. (2020) argue that NLP fairness research lacks real-world grounding. Accordingly, our work responds to this call by examining stereotyping carried through pragmatic mechanisms such as presupposition, implicature, and speech-act framing rather than just overt word choices.

**Pragmatic reasoning in LLMs.** A growing body of work evaluates LLM pragmatic competence. Chen and Wang (2025) propose pragmatic inference chains for improving LLM reasoning on implicit toxic language. Cho and Kim (2024) evaluate LLMs on scalar implicature inference. Yue et al. (2024) test whether LLMs understand conversational implicature through situated dialogue. Mustafin (2025) assesses implicit meaning interpretation in sentiment models from a pragmatic perspective. A comprehensive survey by Mao et al. (2024) confirms that the field remains evaluation-focused, with most work documenting where LLMs fail on pragmatic tasks rather than systematically diagnosing why. Two structural limitations persist across this literature. First, evaluation is overwhelmingly English-centric as cross-linguistic pragmatic evaluation remains rare, and the few non-English studies typically examine a single pragmatic phenomenon in isolation rather than modelling the interaction of multiple mechanisms. Second, existing work treats pragmatic

competence as a general reasoning capacity rather than examining how culture-specific priors shape pragmatic inference; consequently, the question is framed as "can LLMs do pragmatics?" rather than "whose pragmatics do LLMs default to?"

**Arabic NLP and social meaning.** Arabic sentiment analysis has progressed from MSA-focused polarity classification to dialectal and aspect-based approaches (Al-Ayyoub et al., 2019; Abu Farha and Magdy, 2021), but the pragmatic dimension, how (im)politeness, social register, and cultural norms modulate sentiment expression, remains largely unaddressed computationally. Work on Arabic sarcasm detection (Abu Farha and Magdy, 2020; Abuein et al., 2024) treats sarcasm as a classification target rather than examining the cultural-pragmatic mechanisms that produce it. To our knowledge, no existing Arabic NLP resource jointly encodes sentiment and (im)politeness, and no benchmark evaluates LLMs on the interaction between the two in dialectal Arabic.

**The gap.** This paper sits at the intersection of all three dimensions. We provide the first evaluation of LLMs on pragmatically-encoded stereotyping in a non-English language, using a framework that models the interaction of multiple pragmatic mechanisms (implicature, presupposition, mock (im)politeness) grounded in Egyptian Arabic cultural norms. While existing pragmatic evaluation asks whether LLMs can reason pragmatically, we ask whose cultural-pragmatic baseline they reason from, and what happens when that baseline is wrong.

## 3 Pragmatic Stereotyping: Framework and Taxonomy

### 3.1 Pragmatic Stereotyping and Pragmatic Misrecognition

As defined in §1, pragmatic stereotyping is bias conveyed through pragmatic mechanisms rather than explicit lexical content. In our evaluation, we observe two related but distinct phenomena:

**Author-side pragmatic stereotyping (IP):** The commenter encodes prejudice through positive-sentiment surface forms that presuppose negative group attributes. The stereotype lives in the comment; the failure mode is under-detection, as LLMs read the positive surface as playful teasing and miss the prejudicial frame.

**Model-side pragmatic misrecognition (PN, MIP):** The LLM imposes English-trained pragmatic priors on Egyptian Arabic (im)politeness, a register system that operates separately from sentiment. The failure is bidirectional: sincere, polite criticism (PN) is mistrusted as sarcastic (misclassified as MPN), and positive-intended impoliteness (MIP) is read as conflictive. The misrecognition lives in the model's prior, not in the comment; the LLM misrecognizes the (im)politeness register itself.

A single mechanism, insufficient cultural grounding, produces both phenomena.

| Tag | Full Name | S. | (Im)p. | Role |
|---|---|---|---|---|
| PP | Polite Pos. | + | Pol. | Base |
| PoN | Politic Neut. | 0 | Pol-c | Base |
| IN | Impolite Neg. | − | Impol. | Base |
| IP | Impolite Pos. | + | Impol. | Auth. |
| PN | Polite Neg. | − | Pol. | Model |
| MPN | Mock Pol. Neg. | − | M-pol. | Model |
| MIP | Mock Impol. Pos. | + | M-impol. | Model |

Table 1: Seven-tag taxonomy. S. = Sentiment; (Im)p. = (Im)politeness. Base = adequately handled by LLMs; Auth. = author-side stereotyping under-detection; Model = (im)politeness misrecognition. Tags above the mid-rule are explicit (F1 = 0.78–0.85); below are implicit (F1 = 0.62–0.69).

### 3.2 The 7-Tag Annotation Framework

The taxonomy intersects two pragmatic dimensions: sentiment (positive/negative/neutral) and (im)politeness (polite/impolite/politic/mock-polite/mock-impolite). Seven cells are theoretically and empirically populated in Egyptian Arabic discourse (Table 1).

Three tags, PP (Polite Positive), PoN (Politic Neutral), and IN (Impolite Negative), represent cases where surface form and social meaning align; LLMs handle them adequately (F1 = 0.78–0.85). The four tags central to this paper occupy positions in which surface and meaning diverge:

**IP (Impolite Positive):** author-side pragmatic stereotyping. Prejudicial fram-

ing wrapped in positive sentiment, backhanded compliments, grudging praise structured around stereotype, "compliments" that presuppose negative group attributes. The surface form is positive while the pragmatic content is prejudicial.

**PN (Polite Negative):** sincere polite criticism. Genuine disagreement or negative evaluation delivered through mitigated, face-respecting language. The polite surface is sincere, not ironic, but LLMs, lacking this cultural baseline, systematically misread PN as sarcasm (MPN).

**MPN (Mock Polite Negative):** culturally-encoded sarcasm. Surface-polite forms — religious expressions (ما شاء الله), formulaic praise (كلك ذوق), conventional politeness — deployed sarcastically to deliver negative evaluation.

**MIP (Mock Impolite Positive):** positive-intended impoliteness. Impolite linguistic surface (e.g., عيل, يخربيت سنينك مسخرة) deployed in service of a positive social goal: humor, bonding, playful provocation. Rooted in the cultural baseline "المصري ابن نكتة," "Egyptians are born jokers" (Amin, 1953; Al-Tonsi, 2013). Egyptian Arabic normalizes mock-impoliteness as a register of warmth rather than aggression.

### 3.3 The Five-Layer Chain-of-Thought Framework

To make pragmatic reasoning traceable, the taxonomy is operationalized through a five-layer Chain-of-Thought (SI-CoT) framework that decomposes the inference path from comment to tag (Figure 1). Layer 1 identifies context and speech act. Layer 2 distinguishes literal from pragmatic meaning. Layer 3 records modifiers (intensifiers, downtoners, emojis, religious or cultural expressions). Layer 4 infers social goal and sentiment. Layer 5 assigns the (im)politeness judgment and synthesizes the preceding layers into the final tag. The framework draws on speech act theory (Searle, 1969, 1983), Gricean implicature (Grice, 1975), Relevance Theory (Sperber and Wilson, 1995), Brown and Levinson's (1987) politeness model, Culpeper's (2011) impoliteness framework, and Watts's (2003) notion of politic behaviour.

Layers 1–3 operate primarily on observable surface form: speech-act identification, the literal/pragmatic distinction, and modifier extraction can each be supported by lexical and orthographic cues. Layers 4–5 require culture-specific baselines: the social goal of an utterance and its (im)politeness judgment cannot be answered from surface form alone. The framework's diagnostic value lies not in final-tag prediction but in the layer at which reasoning fails, localizing whether failure is lexical or cultural-pragmatic.

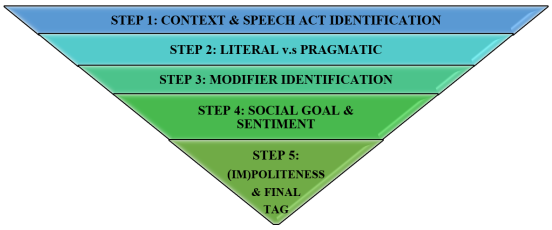


Figure 1: The five-layer SI-CoT annotation framework.

### 3.4 Framework Validity

Two design properties confirm that the taxonomy reflects culturally-grounded distinctions rather than theoretical artefacts. First, expert human annotators applying the SI-CoT guidelines achieve substantial inter-annotator agreement (Cohen's $\kappa = 0.78$) on the seven-tag task across 5,000 Egyptian Arabic social media comments, establishing the benchmark against which LLM performance is measured. Second, an online questionnaire presenting 20 comments to 200 native Egyptian Arabic speakers without linguistic training yielded 67% alignment with expert annotations, well above the 14.3% expected by chance for a seven-way task. The framework encodes distinctions native speakers already make.

## 4 Experimental Setup

### 4.1 Data

The evaluation corpus (NAJAT25) comprises 500 comments drawn by stratified random sampling (random_state=42) from a 5,000-comment gold standard dataset of Egyptian Arabic social media discourse. The corpus and annotation framework are described in full in Assem (2025). The gold standard was collected from Facebook (46%),

TikTok (33%), and Instagram (21%) across six content domains (entertainment, sports, news/politics, food/lifestyle, comedy/memes, social issues/religion), spanning comments from 2023–2024. All comments are 2–20 words in Egyptian Arabic dialect, collected from audio-visual content threads.

Two expert annotators applied the seven-tag SI-CoT framework on the INCEpTION platform, achieving substantial inter-annotator agreement (Cohen's $\kappa = 0.78$) across the full 5,000 comments. Disagreements (22%, n=1,100) were resolved through a three-tier adjudication protocol. The stratified 500-comment test set preserves the gold standard distribution: PP=125 (25.0%), IN=109 (21.8%), PoN=89 (17.8%), PN=54 (10.8%), MPN=49 (9.8%), MIP=42 (8.4%), IP=32 (6.4%). Explicit tags (PP, IN, PoN) and implicit tags (PN, MPN, MIP, IP) are approximately 65% and 35%.

### 4.2 Models and Conditions

Two models were evaluated: GPT-4 (OpenAI) and Claude 3.5 Sonnet (Anthropic), selected as the highest-performing generally-available LLMs at the time of evaluation. Both were tested under two prompting conditions in a within-subjects design, all 500 comments annotated by both models under both conditions, producing four experimental runs.

**Zero-shot.** Models received the full annotation guidelines specifying the seven-tag definitions, the five-layer SI-CoT framework, and the output format. No annotated examples were provided. This condition measures what pragmatic competence LLMs bring from pretraining alone.

**Few-shot.** Models received the same guidelines plus 35 annotated examples (5 per tag), each with complete five-layer CoT reasoning in JSON format. Examples were selected to demonstrate boundary cases, particularly the PN/MPN and MIP/IP distinctions. This condition measures the extent to which in-context examples can substitute for cultural-pragmatic priors.

Both conditions used temperature 0.0 for reproducibility and max_tokens 1,500 to accommodate full CoT output. Comments were processed in randomized batches of 50 (seed=42). Each prompt included the target comment, its video-context metadata (platform, content domain, video topic), and the instruction to produce a structured five-layer CoT annotation followed by a final tag, confidence level, and justification.

### 4.3 Evaluation Metrics

Performance is evaluated at two levels. **Tag-level:** accuracy, Cohen's $\kappa$, and macro-F1. **Diagnostic-level:** per-tag F1 scores to identify the three-tier performance hierarchy, confusion patterns to reveal systematic misclassification directions, and qualitative CoT trace analysis to localize the SI-CoT layer at which reasoning fails. Statistical significance of condition and model effects is assessed via McNemar's test on paired predictions.

## 5 Results and Analysis

All results reported in this section use Claude 3.5 Sonnet few-shot as the primary analysis lens, with GPT-4 few-shot as confirmation. Distribution details are reported in §4.

### 5.1 Overall Performance and the Three-Tier Hierarchy

Both LLMs achieve moderate overall performance under few-shot prompting. Claude reaches 73.6% accuracy (Macro-F1 = 0.72) and GPT-4 reaches 71.4% (Macro-F1 = 0.69). Few-shot prompting shows substantial improvement over zero-shot for both models (+13.2pp for GPT-4, +11.8pp for Claude; McNemar's $\chi^2 > 41$, $p < 0.001$ for both).

Per-tag F1 reveals a three-tier hierarchy (Table 2). High performers (F1 = 0.78–0.85): the explicit tags PP, IN, and PoN, where surface form and social meaning align. Moderate performers (F1 = 0.67–0.69): the implicit tags PN, MIP, and IP, which require pragmatic inference. Low performer: MPN (F1 = 0.62). The discriminating variable is form–meaning alignment, not class frequency or sentiment polarity; the smallest cell (IP, n=32) and the largest implicit cell (PN, n=54) show similar F1, while IN (n=109) substantially outperforms both.

The aggregate gap between explicit-tag F1 (0.81) and implicit-tag F1 (0.66), a 15-percentage-point difference, quantifies the cultural grounding bottleneck at the task level. Few-shot prompting improves all tags, but

| | PP | PoN | IN | PN | MPN | MIP | IP | Mac. |
|---|---|---|---|---|---|---|---|---|
| Claude | .85 | .78 | .79 | .67 | .62 | .67 | .69 | .72 |
| GPT-4 | .83 | .76 | .77 | .64 | .57 | .65 | .67 | .69 |

Table 2: Per-tag F1 scores (few-shot). Mac. = Macro-F1.

with marked asymmetry: implicit tags gain +0.15 to +0.19, while explicit tags gain only +0.08 to +0.10. MPN shows the largest improvement (+0.19) yet remains the lowest performer. In-context examples narrow the gap on culturally-loaded tags, but cannot close it.

### 5.2 Where LLMs Succeed: Evidence Aggregation Without Cultural Inference

Qualitative analysis of Claude's 368 correct few-shot annotations identifies four success patterns, with each comment assigned to its primary driver: (1) explicit linguistic markers aligned with pragmatic intent (n=142, 38.6%), where direct cues reliably produced correct classifications; (2) few-shot example matching (n=89, 24.2%), where structural resemblance to in-context examples drove accuracy even on pragmatically complex cases; (3) strong contextual alignment via metadata (n=78, 21.2%), where convergence between video-genre metadata and internal comment cues made even MIP cases recoverable; and (4) multiple converging cues (n=59, 16.0%), where context, lexicon, emoji, and pragmatic incongruence all aligned. The structural property uniting all four: LLMs succeed when meaning is recoverable from observable input features. The model performs as an evidence aggregator rather than a pragmatic reasoner.

### 5.3 Failure 1: Under-Detection of Author-Side Pragmatic Stereotyping (IP)

The dominant confusion pattern for IP is misclassification as MIP — LLMs read prejudicial-positive content as playful teasing at substantially higher rates than the reverse. IP achieves only 66% correct classification, with MIP representing the single largest destination for misclassified IP comments.

The following example illustrates the mechanism (additional examples in Appendix A):

(1) "محترمة و عارفة ربنا ماشاء الله رغم أنها بشعرها" — "Respectful and God-fearing, mashallah, despite being uncovered." (Intra-Muslim, women's appearance norms.)

Gold-labelled IP, misclassified as MIP. The pragmatic mechanism is concessive presupposition: رغم presents the praised quality as unexpected given the demographic membership. The individual is exempted; the stereotype is reinforced.

Walking the CoT trace makes the failure visible. Layers 1–3 are handled correctly: speech act = praise, surface meaning = positive evaluation, modifiers recorded. The error enters at Layer 4, where social goal is read as convivial and sentiment as positive. The presupposition encoded in رغم انها بشعرها — that uncovered women are normally not respectful or God-fearing — is not recovered. The same Layer 4 failure recurs across distinct stereotype targets (colorism, religious appearance norms), confirming the mechanism is structural rather than target-specific.

The deployment consequence is direct: in a content-moderation pipeline, IP comments classified as MIP pass through as benign banter. The harm falls on the demographic groups targeted by pragmatic stereotyping, and the harm is silent.

### 5.4 Failure 2: Bidirectional Reversal of (Im)politeness Reading (PN, MIP)

The second failure manifests in two opposite directions through one mechanism: LLMs cannot read Egyptian Arabic (im)politeness as a register-system that operates separately from sentiment.

#### 5.4.1 The Polite Side: Sincere Politeness Misread as Sarcasm

PN comments are systematically misclassified as MPN at rates far exceeding the reverse direction. MPN itself is the lowest-performing tag (F1 = 0.62), and high-confidence MPN predictions are correct only 68% of the time — the model is not merely wrong but systematically overconfident in the wrong reading. The following example illustrates the pattern (additional examples in Appendix A):

(4) "مع احترامي لرأيك بس الكلام ده مش دقيق في تفاصيل

كتير ناقصة" — "With respect to your opinion, but this isn't accurate.. there are many missing details."

Gold-labelled PN. The structural property is concessive politeness: a polite preamble followed by softened criticism, with no irony markers. The failure originates at Layer 5 — the (im)politeness judgment — where the model cannot represent sincere polite criticism as a coherent speech-act category. Its English-internet-derived prior treats polite-looking online speech that is not transparently positive as probably ironic.

#### 5.4.2 The Impolite Side: Positive-Intended Impoliteness Misread as Conflictive

MIP comments are misclassified as hostile (IP or IN) at non-trivial rates, with MIP achieving only 67% correct classification. High-confidence MIP predictions are correct only 71% of the time. The following example illustrates the pattern (additional examples in Appendix A):

(6) "يخربيت سنينك" (with positive emojis) — "May your years be ruined" — conventionalized playful insult.

Gold-labelled MIP. The CoT trace localizes the failure at Layer 5: the model correctly registers the situational context as comedic at Layer 1 and indexes impolite surface forms at Layers 2–3, yet at Layer 5 the (im)politeness judgment flips to genuinely impolite. The genre metadata supplied at input does not propagate into the Layer 5 inference.

#### 5.4.3 One Mechanism, Two Reversals

The polite-side and impolite-side failures are instances of the same gap: the LLM lacks a prior for Egyptian Arabic (im)politeness as a register-system in which surface form can carry the opposite of its expected social meaning. Sincere politeness triggers a fake-politeness reading. Positive-intended impoliteness triggers a genuine-impoliteness reading. The same missing baseline produces opposite reversals — over-attribution of sarcasm to polite forms and over-attribution of hostility to playful impolite forms.

### 5.5 The Evidence-Aggregation Axis

LLMs perform reliably when meaning is recoverable from observable surface features and fail systematically when meaning requires inference from culture-specific priors. The 15-percentage-point F1 gap between explicit tags (0.81) and implicit tags (0.66) quantifies this gradient. Few-shot prompting partially closes the gap but cannot close it entirely: in-context examples substitute for cultural priors when test comments structurally resemble the examples, but not when novel comments require genuine inference from priors the model lacks.

## 6 Conclusion

This paper introduced the concept of *pragmatic stereotyping* — stereotyping conveyed through pragmatic mechanisms rather than explicit lexical content — and demonstrated that cultural grounding is the critical bottleneck preventing LLMs from detecting it in non-English discourse. Evaluating GPT-4 and Claude 3.5 Sonnet on 500 Egyptian Arabic social media comments, we found that approximately 35% of LLM errors stem from cultural grounding gaps, producing a 15-percentage-point F1 gap between explicit and implicit pragmatic tags.

These findings address three gaps: the diagnostic deficit identified by Mao et al.'s (2024) survey of pragmatic processing, the cultural bias embedded in English-centric pragmatic inference approaches such as Chen and Wang (2025) and Cho and Kim (2024) (which assume English pragmatic norms), and Blodgett et al.'s (2020) call for fairness research grounded in how bias actually operates in natural language.

Pragmatic stereotyping lives in what is meant, and detecting it requires cultural-pragmatic competence that current LLMs lack for non-English varieties. Future work should extend this evaluation to other Arabic dialects and other non-English languages whose pragmatic systems diverge from English norms, and investigate whether cultural knowledge integration can narrow the gap.

## Limitations

This study evaluates two LLMs on one Arabic dialect using a single dataset; findings may

not generalize to other Arabic varieties, languages, or model versions. The test set yields small cell sizes for low-frequency tags (IP=32), and the SI-CoT framework may require adaptation for other registers. Our evaluation uses the models' own CoT outputs as diagnostic evidence, which may not reflect actual internal reasoning.

## Ethics Statement

The dataset consists of publicly posted social media comments collected in accordance with platform terms of service. All comments are anonymized: usernames, profile information, and identifying metadata are removed. The annotation framework was developed with attention to the cultural positionality of the annotators as native Egyptian Arabic speakers, and the native-speaker validation study was conducted with informed consent. The examples of pragmatic stereotyping presented in this paper (sexism, colorism, religious prejudice) are reproduced for analytical purposes; their inclusion does not constitute endorsement.

## References


Ibrahim Abu Farha and Walid Magdy. 2020. From Arabic sentiment analysis to sarcasm detection: The ArSarcasm dataset. In *Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools*, pages 32–39. European Language Resource Association.

Ibrahim Abu Farha and Walid Magdy. 2021. Benchmarking transformer-based language models for Arabic sentiment and sarcasm detection. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 21–31. Association for Computational Linguistics.

Q. Abuein, R. M. Al-Khatib, A. Migdady, M. S. Jawarneh, and A. Al-Khateeb. 2024. ArSa-tweets: A novel Arabic sarcasm detection system based on deep learning model. *Heliyon*, 10(17):e36892.

Mahmoud Al-Ayyoub, Abed Allah Khamaiseh, Yaser Jararweh, and Mohammed N. Al-Kabi. 2019. A comprehensive survey of Arabic sentiment analysis. *Information Processing & Management*, 56(2):320–342.

Abbas Al-Tonsi. 2013. *Umm Al-Dunya: Advanced Egyptian Colloquial Arabic*. The American University in Cairo Press.

Ahmed Amin. 1953. *Dictionary of Egyptian Customs, Traditions and Expressions*. Hindawi Foundation. Reprinted 2013.

Samar Assem. 2025. *Building an Annotated Corpus for Egyptian Arabic Sentiment Analysis: A Computational Linguistics Approach*. Ph.D. thesis, Alexandria University.

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. Association for Computational Linguistics.

Penelope Brown and Stephen C. Levinson. 1987. *Politeness: Some Universals in Language Usage*. Cambridge University Press.

X. Chen and S. Wang. 2025. Pragmatic inference chain (PIC): Improving LLMs' reasoning of authentic implicit toxic language. *arXiv*.

Y. Cho and S. Kim. 2024. Pragmatic inference of scalar implicature by LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 10–20. Association for Computational Linguistics.

Jonathan Culpeper. 2011. Politeness and impoliteness. In Wolfram Bublitz, Andreas H. Jucker, and Klaus P. Schneider, editors, *Pragmatics of Society*, volume 5, pages 393–438. De Gruyter Mouton.

H. Paul Grice. 1975. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics 3: Speech Acts*, pages 41–58. Academic Press, New York.

R. Mao, M. Ge, S. Han, W. Li, K. He, L. Zhu, and E. Cambria. 2024. A survey on pragmatic processing techniques. *Information Fusion*, 114:102712.

R. Mustafin. 2025. Pragmatic perspective on assessing implicit meaning interpretation in sentiment analysis models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 898–907. Association for Computational Linguistics.

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*, pages 5356–5371. Association for Computational Linguistics.

Nikita Nangia, Clara Vania, Rajesh Bhatt, 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*, pages 1953–1967. Association for Computational Linguistics.

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

John R. Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge University Press.

John R. Searle. 1983. *Intentionality: An Essay in the Philosophy of Mind*. Cambridge University Press.

Dan Sperber and Deirdre Wilson. 1995. *Relevance: Communication and Cognition*, 2nd edition. Blackwell, Oxford.

Richard J. Watts. 2003. *Politeness*. Cambridge University Press, Cambridge.

S. Yue, S. Song, X. Cheng, and H. Hu. 2024. Do large language models understand conversational implicature: A case study with a Chinese sitcom. *arXiv*.

# A Additional Example Comments

## A.1 Author-Side Pragmatic Stereotyping (IP → MIP misclassification)

(2)"ملتحي بس مش متزمت خالص" — "Bearded, but not extremist at all." (Intra-Muslim, men's appearance and extremism stereotype.) Gold-labelled IP, misclassified as MIP. The concessive "بس" presents the absence of extremism as unexpected given the beard, presupposing that bearded men are normally extremist.
(3) "سمرة بس وشها زي القمر" — "Dark-skinned, but her face is like the moon." (Colorism.) Gold-labelled IP, misclassified as MIP. The concessive بس presents beauty as unexpected given dark skin, presupposing that dark-skinned women are normally not beautiful.

## A.2 Sincere Politeness Misread as Sarcasm (PN → MPN misclassification)

(5) ممكن الأسلوب ده يكون قاسي شوية على الأطفال مش كده؟ [thinking face emoji]"— "Could this approach be a bit harsh on children, don't you think? " (Pedagogical register.) Gold-labelled PN, misclassified as MPN. The thinking emoji contextually signals reflection rather than mockery; the model overrides this counter-evidence in favor of a sarcastic reading.

## A.3 Positive-Intended Impoliteness Misread as Conflictive (MIP misclassification)

(7) "عيل مسخرة" — "You ridiculous kid" — diminutive-mockery deployed as affectionate ribbing. Gold-labelled MIP. The model reads the impolite surface as genuinely hostile despite comedic context.

## A.4 Success Pattern Examples

"أجمد كدة مفيش رجالة بتعيط كله هيعدي"— "That's the spirit, real men don't cry, this'll pass." Correctly classified as IP through few-shot example matching: its toxic-masculinity-as-encouragement structure matched the IP in-context examples.
"حرام عليك هترفد بسببك همووووت"— "Shame on you, you'll get me fired, I'm dying [laughing]." In a comedy TikTok thread, correctly classified as MIP through convergence of comedy-genre

metadata, playful exaggeration, and the elongated expressive-laughter marker (همموووت).

## B Prompt Template

The following prompt was used in the few-shot condition. In the zero-shot condition, the same prompt was used without the annotated examples block. The few-shot condition included 35 examples (5 per tag), each with complete five-layer CoT output in JSON format.The full set of 35 few-shot examples is available upon request.

```
You are an expert annotator specializing
in Egyptian Arabic social media comments.
Your task: Assign ONE of 7 pragmatic-
sentiment tags to the comment below.

TAGS:
- PP: Polite Positive
- MIP: Mock Impolite Positive
- IP: Impolite Positive
- PoN: Politic Neutral
- IN: Impolite Negative
- MPN: Mock Polite Negative
- PN: Polite Negative

OUTPUT SCHEMA (JSON):
{
 "annotation": {
  "step_1_context_speech_act": {
   "situational_context": "...",
   "main_speech_act": "..."
  },
  "step_2_literal_pragmatic": {
   "literal_meaning": "...",
   "explicitness": "Explicit|Implicit",
   "pragmatic_interpretation": "..."
  },
  "step_3_modifiers": {
   "internal_modifiers": ["..."],
   "external_modifiers": ["..."]
  },
  "step_4_social_goal_sentiment": {
   "social_goal": "Convivial|Conflictive
                  |Neutral",
   "sentiment": "Positive|Negative
                |Neutral"
  },
  "step_5_politeness_tag": {
   "politeness_assessment": "...",
   "sincerity": "Sincere|Insincere",
   "final_tag": "PP|MIP|IP|PoN|IN
                |MPN|PN",
   "confidence": "High|Medium|Low",
   "justification": "..."
  }
 }
}

COMMENT TO ANNOTATE: "{comment}"
```

## C Full Per-Tag F1 Results

| | PP | PoN | IN | PN | MPN | MIP | IP | Mac. |
|---|---|---|---|---|---|---|---|---|
| GPT-4 (ZS) | .74 | .65 | .66 | .45 | .36 | .48 | .51 | .55 |
| GPT-4 (FS) | .83 | .76 | .77 | .64 | .57 | .65 | .67 | .69 |
| Claude (ZS) | .77 | .69 | .69 | .51 | .43 | .51 | .54 | .59 |
| Claude (FS) | .85 | .78 | .79 | .67 | .62 | .67 | .69 | .72 |

Table 3: Per-tag F1 across all conditions. ZS = zero-shot; FS = few-shot; Mac. = Macro-F1.