

Arabic Author Gender Profiling with Large Language Models: An Empirical Investigation of Capability, Calibration, Bias, and Prompt Sensitivity

Tarfah Alrashid

`ttal-rashid@uj.edu.sa`

University of Jeddah

Research Article

Keywords: Arabic NLP, Author Profiling, Large Language Models, Gender Bias, Prompt Sensitivity, PAN 2018

Posted Date: October 16th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10307505/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Arabic Author Gender Profiling with Large Language Models: An Empirical Investigation of Capability, Calibration, Bias, and Prompt Sensitivity

Tarfah Alrashid^{1*}

^{1*}Department of Computer Science and Artificial Intelligence,
University of Jeddah, Jeddah, Saudi Arabia .

Corresponding author(s). E-mail(s): ttal-rashid@uj.edu.sa;

Abstract

Author gender profiling, with applications in forensics and security, has been transformed for English by large language models (LLMs), which surpass supervised shared-task systems zero-shot. Whether this advantage transfers to Arabic, a morphologically rich language spoken by over 400 million people, is untested: prior LLM contact with Arabic gender inference is limited to a preliminary multimodal probe and to gender prediction from profile metadata. We present, to the best of our knowledge, the first comprehensive study of text-based LLM author gender profiling in Arabic, evaluating Gemini 2.5 Flash, GPT-5.4 mini, and the Arabic-native Jais-2-8B-Chat on the official PAN 2018 Arabic test partition (1,000 authors) across two text-volume levels, two prompt variants, and two prompt languages. Four findings emerge. First, the English advantage does not transfer: all three LLMs fall below the best classical system of the 2018 shared task (81.70%), the multilingual models by 3.2–3.5 points, the Arabic-native model by 13.8. Second, abstention behaviour is well calibrated: abstention rises from under 2% at author level to 83.9% at tweet level, and forcing predictions on abstained content collapses accuracy to near chance, making abstention a usable reliability signal. Third, a female-leaning bias appears across all models under English prompting but vanishes under forced binary prompting, showing that bias is confidence-modulated rather than fixed. Fourth, Arabic prompts reverse the Arabic-native model’s bias direction from female-leaning to male-leaning, a previously undocumented pattern. Error analysis identifies an Arabic-specific failure mode in which addressee-gender morphology is misread as author gender, and a preprocessing mitigation.

Keywords: Arabic NLP, Author Profiling, Large Language Models, Gender Bias, Prompt Sensitivity, PAN 2018

1 Introduction

Author profiling, the task of inferring demographic and psychological attributes from written text, has broad applications in cybersecurity, marketing, forensic linguistics, and personalised NLP systems [1, 2]. Gender identification, one of the most studied author profiling subtasks, has historically been approached through feature engineering and classical machine learning, achieving strong results on well-established benchmarks such as the PAN shared task series [1, 3]. The emergence of transformer-based large language models (LLMs) [4] raises fundamental questions that go beyond simple performance comparison: do these general-purpose systems behave reliably on Arabic text? Do they exhibit systematic gender bias, and if so, under what conditions? What Arabic-specific linguistic mechanisms explain their failures? These questions have direct implications for the deployment of LLMs in Arabic NLP systems and for our understanding of cross-lingual LLM behaviour.

This question has been partially answered for English. Staab et al. [5] showed that GPT-4 achieves 90.2% accuracy on PAN 2018 English gender classification in a zero-shot setting, surpassing the best competition system (82.21%) by approximately 8 points. However, their study was explicitly restricted to English, leaving the Arabic subtask, the most morphologically complex of the four benchmark languages, entirely unevaluated. This gap is significant: Arabic is spoken by over 400 million people, is morphologically complex, and has been consistently shown to be more challenging for LLMs than English across a wide range of NLP tasks [6–8]. Whether the dramatic LLM advantage observed in English translates to Arabic, or whether the English-Arabic capability gap undermines it, remains an open empirical question.

Beyond performance, the use of LLMs for gender classification raises important fairness concerns. LLMs trained on web-scale text inherit the gender stereotypes present in their pretraining data, and these stereotypes may systematically bias classification outcomes in ways that disadvantage certain authors [9, 10]. For Arabic specifically, the language’s rich grammatical gender morphology creates unique opportunities for LLM misinterpretation (for example, confusing addressee-gender markers with author-gender markers) that do not arise in English. The characterisation of these bias mechanisms in Arabic LLMs has only recently begun to receive attention [11], and remains unexplored for author-attribute inference.

A further underexplored dimension concerns the effect of prompt language. Prior work has established that English prompts consistently outperform Arabic prompts for multilingual LLMs on Arabic NLP tasks [6, 7, 12], suggesting that English should be the default prompt language even for Arabic tasks. This advantage extends, in aggregate accuracy terms, even to Arabic-centric models [6]; however, how prompt language shapes an *Arabic-native* model’s behaviour, its coverage, calibration, and bias, has not been examined. If Arabic-native training reverses the English-prompt

advantage, practitioners would face a fundamentally different prompt design decision when deploying such models.

This paper addresses these questions through a systematic empirical study of three LLMs on the PAN 2018 Arabic gender classification dataset: Gemini 2.5 Flash, GPT-5.4 mini, and Jais-2-8B-Chat. We study three state-of-the-art LLMs under multiple experimental conditions designed to examine performance, calibration, prompt sensitivity, and gender bias: two text granularity levels (author-level with 100 tweets and tweet-level with a single tweet), two prompt variants (standard allowing abstention and forced binary), and two prompt languages for the Arabic-native model (English and Arabic). This paper addresses four interrelated research questions:

1. **Cross-lingual capability:** Does the LLM advantage over classical methods documented for English author gender profiling transfer to Arabic?
2. **Calibration and reliability:** Do LLMs accurately self-assess their classification uncertainty on Arabic text, and what content properties govern abstention behaviour?
3. **Fairness and bias:** Do LLMs exhibit systematic gender bias on Arabic text, and how is this bias modulated by prompt design, text volume, and confidence?
4. **Arabic-specific linguistic analysis:** What Arabic-specific linguistic mechanisms explain LLM failures, and how does prompt language affect an Arabic-native LLM’s behaviour and bias direction?

To the best of our knowledge, this is the first comprehensive study of text-based LLM behaviour on the official PAN 2018 Arabic author-profiling test partition, prior LLM contact with this benchmark’s Arabic portion being limited to a preliminary multimodal probe [13], and no previous work examines performance, calibration, bias, and prompt sensitivity simultaneously. Within this framework, our main contributions are:

1. **An English–Arabic LLM capability gap.** All three LLMs fall below the best classical SVM system of the 2018 shared task, in stark contrast to the English setting, where zero-shot LLMs substantially exceed the supervised baseline [5]. Zero-shot LLM deployment for Arabic author profiling cannot assume English-level performance.
2. **A text-volume threshold governing calibration and abstention.** Abstention rises sharply from 0–1.5% at author level to 47.3–83.9% at tweet level, and forced predictions on previously-abstained tweets collapse to near-chance accuracy (60.88%), showing that LLMs accurately self-assess classification difficulty and that abstentions reflect genuine uncertainty rather than over-caution.
3. **Confidence-modulated gender bias.** A consistent female-leaning bias appears across all three models under English prompting, with recall gaps ranging from 7 to 42 percentage points, and disappears entirely under forced binary prompting for Gemini and GPT-5.4 mini: gender bias in Arabic LLMs is linked to model confidence mechanisms rather than being a fixed classification property.
4. **A prompt-language-dependent bias reversal in an Arabic-native LLM.** We present the first behavioural analysis of how prompt language affects an Arabic-native model, examining coverage, calibration, and bias direction beyond

the aggregate accuracy comparisons of prior work. For Jais-2-8B-Chat, Arabic prompts improve precision but reduce coverage and reverse the bias direction from female-leaning to male-leaning (38-point gap), with the accuracy advantage statistically non-significant and attributable to selective abstention rather than genuine capability improvement.

5. **Arabic-specific failure modes.** Qualitative error analysis identifies three misclassification profiles: religious content ambiguity, Arabic morphological confusion, and stereotypical topic association. The morphological profile, in which addressee-gender markers are misread as author-gender signals, is linguistically specific to Arabic and has no English analogue; a stereotype override phenomenon, in which gender priors overrule the model’s own explicit reasoning, is documented for the first time in Arabic.

The remainder of this paper is organised as follows. Section 2 reviews related work on author profiling, Arabic LLM behaviour, gender bias, and prompt language effects. Section 3 describes the dataset, models, experimental design, and evaluation metrics. Section 4 presents all experimental results. Section 5 synthesises findings and discusses implications. Section 6 concludes with limitations and future work directions.

2 Related Work

2.1 Author Profiling and Gender Classification

Author profiling, the task of inferring demographic and psychological attributes of authors from their writing style, counts gender identification among its most extensively studied subtasks [1, 14]. Early work showed that Twitter text carries sufficient stylistic and lexical signal for gender inference, with n-gram and sociolinguistic features reaching around 70% accuracy [14], and Burger et al. [15] confirmed reliable gender signals at scale over 180,000 accounts; Nguyen et al. [16] situate the task within the broader computational sociolinguistics literature. The PAN shared task series at CLEF standardised evaluation across languages, with the 2017 and 2018 editions addressing Twitter-based gender identification in Arabic among four target languages [1, 3]. Feature-engineering approaches dominated both editions: the best Arabic gender result reached 80.31% at PAN 2017 [3], while at PAN 2018 the best Arabic text-only result, 81.70%, was achieved by Tellez et al. [17] with an SVM over word and character n-grams and skip-grams, and Daneshvar and Inkpen [18] recorded the best cross-language textual average, including the 82.21% English figure that Staab et al. [5] later adopted as their reference point.

Neural approaches followed. Zhang and Abdul-Mageed [19] fine-tuned BERT models for Arabic author profiling in the APDA shared task, reaching 81.67% gender accuracy on Twitter data, and Bsir et al. [20] fine-tuned ARBERT, MARBERT, AraBERT, and BERT-base-Arabic with multi-task learning on PAN author-profiling data, reporting a best accuracy of 79.9% with AraBERT, though on a custom 80/20 split rather than the official test partition, which precludes direct comparison with shared-task results. These efforts build on Arabic-specific pretrained encoders such as AraBERT [21] and MARBERT [22], which established strong baselines for Arabic

social media NLP. In a complementary setting, Alshaalan and Al-Khalifa [23] combined deep learning with handcrafted features for gender identification in MSA news text, contributing a large gender-labelled article dataset.

LLM-based gender inference has so far touched Arabic only peripherally. Lyu et al. [13] probe GPT-4V on 500 randomly sampled image-text pairs per PAN 2018 language, reaching 70.0% accuracy on Arabic without baseline comparison, calibration analysis, or the official author-profiling protocol; LAraBench’s gender task operates on Twitter profile metadata (primarily user names) rather than authored text [8]; and two recent preprints work in complementary directions: Alsadhan [24] ask whether Arabic and English text *generated* by LLMs mimics human personality markers, machine mimicry rather than machine inference of human authors’ attributes, and Lafargue et al. [25] probe zero-shot author profiling for cultural signals without addressing Arabic, gender, benchmark comparison, or calibration. No study has systematically evaluated LLMs on text-based Arabic author gender profiling, and the official PAN 2018 Arabic test partition has not been used in a dedicated LLM evaluation — the gap this work fills.

2.2 Large Language Models for Arabic NLP

The behaviour of LLMs on Arabic NLP has become an active research area. Abdelali et al. [8] analysed LLM performance across 33 tasks spanning 61 datasets, finding that task-specific models generally outperform LLMs in zero-shot settings, with few-shot adaptation narrowing the gap for larger models. Koto et al. [6] introduced ArabicMMLU, a native Arabic multitask benchmark on which GPT-4 achieves the highest accuracy (72.5%) while the Arabic-centric Jais leads open-source alternatives. Evaluation activity continues to expand: recent 2026 studies benchmark LLMs on Arabic parsing [26] and evaluate Jais-2-8B-Chat alongside frontier proprietary models on dialectal and cultural dialogue tasks [27].

Arabic-centric open models, of which the Jais family [28, 29] is the most widely used representative, test whether language-focused pretraining can offset the scale and diversity advantages of multilingual frontier systems; we describe Jais-2-8B-Chat in Section 3.2. Yet across this rapidly growing evaluation landscape, author profiling is absent: a systematic survey of more than 40 Arabic LLM benchmarks [30] organises existing efforts into knowledge, NLP-task, cultural/dialectal, and safety-oriented categories, with author profiling tasks such as gender classification featuring in none of them, underscoring the gap our work addresses.

2.3 Gender Bias in Language Models

Gender bias in NLP systems has been extensively documented for English [9, 10, 31]. Kotek et al. [10] showed that all tested LLMs follow gender stereotypes in ambiguous pronoun resolution, consistently pairing stereotypically male occupations with “he” and female occupations with “she”, and Gallegos et al. [9] provide a broad survey of bias sources, evaluation methods, and mitigation strategies whose taxonomy we draw on in our analysis.

Non-English languages have received far less attention [32, 33]. Dedicated Arabic resources are only now emerging: ArGAN evaluates gender, ability, and nationality biases in Arabic LLM outputs [11], and ArabJobs benchmarks gender-bias detection over Arabic job advertisements [34]; both examine bias in model *generation* and static associations rather than in the inference of author attributes. Arabic is a particularly demanding case: grammatical gender is marked on verbs, adjectives, and nouns, producing bias manifestation patterns with no English counterpart. Prior Arabic work on grammatical gender has likewise focused on the generation side, identifying and rewriting first- and second-person gender references to produce correctly gendered text [35]. The inference-side error we document, LLMs misreading addressee-gender morphology as author-gender evidence, has no precedent in this literature and constitutes a previously undocumented Arabic-specific bias mechanism.

2.4 Prompt Language Effects in Arabic NLP

The effect of prompt language on LLM performance on Arabic tasks has been studied across several task types. Khondaker et al. [7] evaluated ChatGPT with both English and Arabic prompts on multiple Arabic NLP tasks including dialect identification, toxicity detection, and machine-generated text detection, finding that English prompts consistently outperformed Arabic counterparts, leading them to adopt English as the standard prompt language for Arabic NLP studies. Kadaoui et al. [12] similarly found a 1–2 BLEU point advantage for English prompts over Arabic prompts in Arabic machine translation experiments, substantiating the English prompt superiority finding across task types.

At a larger scale, Koto et al. [6] benchmarked all open-source models on ArabicMMLU under four prompt-language settings and found that the optimal configuration across every model was an English prompt with English-script answer options. Notably, their ablation included the Arabic-centric Jais-chat and AceGPT-chat, which proved more robust than multilingual models to Arabic-script output, an early indication that Arabic-centric training modulates prompt-language sensitivity, even though the aggregate English-prompt advantage held for these models as well.

Existing comparisons, however, report aggregate accuracy on knowledge benchmarks. None examines how prompt language shapes an Arabic-native model’s *behaviour* on a fixed task: its accuracy–coverage trade-off, abstention, format compliance, and, critically, the direction of its bias. Our prompt-language study on Jais-2-8B-Chat provides the first such behavioural analysis, revealing a nuanced pattern: Arabic prompts improve precision but reduce coverage, and produce an unexpected reversal in gender bias direction, findings with direct implications for practitioners deploying Arabic-native LLMs.

2.5 LLM Calibration and Abstention Behaviour

The ability of LLMs to express uncertainty through abstention is increasingly recognized as an important property for reliable deployment [9, 36]. Well-calibrated models should commit when they have sufficient evidence and abstain when they do

not. Probability-based confidence calibration has been examined for LLMs on ArabicMMLU [6]; *behavioural* calibration expressed through verbalised abstention has not. Our calibration analysis (Section 4.6) provides direct evidence of the latter in the Arabic gender classification setting: forced predictions on previously-abstained content collapse to near-chance accuracy, and abstention rises sharply as available text shrinks, revealing a text-volume threshold effect not previously characterised for Arabic.

3 Methodology

3.1 Dataset

We evaluate on the Arabic subset of the PAN 2018 Author Profiling shared task dataset [1], which consists of 1,000 Twitter author profiles, each comprising 100 tweets. The dataset is perfectly balanced, containing 500 male-authored and 500 female-authored profiles. We use the official PAN 2018 Arabic test partition, which includes gold-standard gender labels released by the shared task organisers [1], in contrast to Staab et al. [5], who evaluated on a same-sized sample of the training set because the official English test set was not available to them. Author identities are anonymised via hashed identifiers, and tweet content is provided in raw XML format. The dataset covers a broad range of Arabic Twitter content, including Modern Standard Arabic (MSA), dialectal Arabic, religious discourse, political commentary, and casual conversation. Following prior work [5], we restrict evaluation to the text modality only, excluding the 10 images per profile also provided in the original dataset.

This dataset has been used as the benchmark for the PAN 2018 Author Profiling shared task, which attracted 23 participating teams. The best-performing text-only system on Arabic, by Tellez et al. [17], achieved 81.70% accuracy using an SVM over n-gram and skip-gram features; the second- and third-best Arabic textual systems reached 81.20% and 80.90% [18], with no statistically significant difference among the top three [1]. We adopt the best result (81.70%) as our primary baseline.

Data contamination considerations. Because the PAN 2018 corpus has been publicly available since 2018, its tweets may overlap with the pretraining data of the evaluated models. Several observations suggest, however, that memorisation is unlikely to drive our results. First, gender labels are distributed only through the shared task’s truth files, in which authors are identified by hashed identifiers; the mapping from tweet content to gender label does not occur in natural web text, making direct label memorisation implausible. Second, memorisation would be expected to inflate performance, whereas all three models fall *below* the 2018 feature-based baseline. Third, the models’ tweet-level behaviour (abstention rates of 47.3–83.9% and near-chance accuracy under forced prediction) is consistent with genuine inference under uncertainty rather than recall of previously seen label associations. Finally, Staab et al. [5] conducted an explicit decontamination study on the English PAN 2018 data and found no evidence that memorisation affected GPT-4’s predictions. Nevertheless, indirect contamination cannot be fully excluded, and we note this in the limitations.

3.2 Models

We study three large language models spanning different training paradigms:

Gemini 2.5 Flash [37] is a general-purpose multilingual model developed by Google DeepMind, accessed via the Google AI Studio API. It supports Arabic among its many languages and features a controllable thinking budget, which we set to zero (`thinking_budget=0`) to ensure full output token allocation for classification responses and to avoid hidden reasoning tokens consuming the available context.

GPT-5.4 mini [38] is a cost-efficient multilingual model developed by OpenAI, accessed via the OpenAI API. It represents the current generation of OpenAI’s compact model family, optimised for instruction-following tasks.

Jais-2-8B-Chat [29] is an Arabic-native open-weight large language model developed by Inception (G42, Abu Dhabi), MBZUAI, and Cerebras, the second generation of the Jais family [28]. Unlike the two multilingual models, Jais-2 was trained from scratch on an Arabic-first bilingual dataset of Arabic and English, with a redesigned architecture and a custom Arabic-centric vocabulary designed to capture Modern Standard Arabic, regional dialects, and Arabic–English code-switching. It is the most widely used Arabic-specialist LLM family in the research community and serves as a critical comparison point for assessing whether Arabic-native training confers an advantage on Arabic NLP tasks. We load Jais-2-8B-Chat in full float16 precision on an NVIDIA GPU via the Hugging Face Transformers library [39], using the GitHub development version of the library as required by the model’s architecture.

3.3 Experimental Design

We examine each model under two text granularity conditions:

Author-level classification follows the original PAN 2018 protocol [1] and the methodology of Staab et al. [5]: all 100 tweets from a single author are concatenated into one prompt, and the model is asked to predict the author’s gender based on the full profile. This setting allows direct comparison against PAN 2018 shared task results.

Tweet-level classification presents the model with a single tweet and asks it to infer the author’s gender based solely on that tweet. Single-tweet gender classification has been examined in the supervised setting since Burger et al. [15], and tweet-level modelling is a standard strategy in transformer-based author profiling [20]; what is novel here is its use as a *zero-shot LLM evaluation condition*: supervised tweet-level classifiers cannot abstain, whereas this condition allows us to measure how LLM accuracy, calibration, and abstention behave under minimal text signal, a question no prior study on this benchmark has addressed. To ensure efficiency and reproducibility, tweets are processed in batches of five per API call. Critically, each batch contains tweets from five *different* authors, randomly shuffled (seed=42) prior to batching to prevent the model from detecting shared authorship within a batch. This design preserves the single-tweet classification condition while reducing the number of API calls from 100,000 to 200. The tweet-to-batch assignment was generated once, saved to persistent storage, and loaded identically for all models, guaranteeing fair cross-model comparison.

Under each text granularity condition, we examine two prompt variants:

Standard prompt instructs the model to predict the author’s gender (Male, Female, or Cannot be determined), explicitly permitting abstention when the model cannot make a confident prediction.

Forced binary prompt instructs the model to predict either Male or Female with no abstention option, even under uncertainty. This variant is used to measure the effect of forcing binary outputs on both accuracy and gender bias.

This study deliberately focuses on zero-shot LLM behaviour, assessing what LLMs can infer without task-specific training data. Fine-tuned transformer baselines such as MARBERT [22] were not included because the objective of this study is to characterise zero-shot LLM behaviour rather than to establish a supervised state of the art: no fine-tuned transformer result has been published on the official PAN 2018 Arabic test partition (the closest study, Bsir et al. [20], reports 79.9% with fine-tuned AraBERT on a custom split of PAN author-profiling data), and training one would answer a different research question. The SVM baselines [17, 18] are retained as the reference points: both are supervised systems, trained on the labelled 1,500-author PAN 2018 Arabic training partition, and thus represent what dedicated task-specific supervised learning achieved on this benchmark. The comparison between zero-shot LLMs and these supervised baselines is therefore deliberately asymmetric; we adopt it because it mirrors exactly the comparison made for English by Staab et al. [5], where zero-shot GPT-4 surpassed the best supervised competition system despite the same asymmetry. Whether zero-shot LLMs can match systems trained specifically for the task is precisely the question of practical interest. The evaluation of fine-tuned Arabic transformer models on this task is left as future work.

3.4 Prompt Design

Our prompting approach combines zero-shot instruction [40] with chain-of-thought reasoning, which Wei et al. [41] showed substantially improves reliability on complex classification tasks. All prompts are adapted from Staab et al. [5], who applied a similar zero-shot guessing game framing to English PAN 2018 data. We extend their design to Arabic and to the batch tweet-level setting. The author-level prompt follows their original structure:

“Let’s play a guessing game. Below are publicly available tweets written by one author in Arabic. Your task is to guess whether the author is male or female. Base your guess on all available information in the text, including writing style, word choice, topics, and any linguistic cues. First, reason step-by-step. Then give your final answer in the format: Reasoning: [...] Answer: [Male/Female]”

For the tweet-level setting, the prompt is modified to present multiple tweets from different authors, explicitly instructing the model to treat each independently. The forced binary variant adds the instruction: *“You must answer either Male or Female for every tweet. Even if you are not certain, make your best guess.”*

All models share an identical system prompt establishing the investigator persona, adapted from Staab et al. [5]: *“You are an expert investigator with experience in online profiling and Arabic text analysis. You work with an analytical mindset and try to answer as precisely as possible.”*

3.5 Prompt Language Study

For our primary experiments with all three models, we use English prompts, following the consistent English-prompt advantage documented on Arabic tasks (Section 2.4) [6, 7, 12].

To examine the effect of prompt language on an Arabic-native LLM, we additionally analyse Jais-2-8B-Chat under fully Arabic-language prompts. In this condition, the system prompt, task description, format instructions, and output labels are all translated to Arabic, while tweet content remains unchanged. Output labels are translated as: Male→*thakar*, Female→*untha*, Cannot be determined→*la yumkin tahdiduh*. This yields four Jais experimental conditions in total: English standard, English forced, Arabic standard, and Arabic forced. As discussed in Section 2.4, no prior study isolates these behavioural effects of prompt language for an Arabic-native model; this comparison provides the first such analysis.

3.6 Inference Parameters

For API-based models (Gemini, GPT-5.4 mini), we set temperature to 0.1 following Staab et al. [5], to ensure near-deterministic outputs while allowing minor variation. Maximum output tokens are set to 600 for author-level prompts and 1,000 for tweet-level batch prompts, to accommodate the model’s reasoning trace before the final answer.

For Jais-2-8B-Chat, we use greedy decoding (`do_sample=False`) following the model’s official inference guidelines, which is equivalent to temperature=0 and ensures fully deterministic outputs. Maximum new tokens are set to 600 for author-level and 800 for tweet-level batch prompts.

3.7 Response Parsing

Model responses are parsed to extract the final gender prediction. Each response is expected to follow the structured format **Answer:** [Male/Female] or **Answer:** [Cannot be determined]. We implement model-specific parsers to handle observed output variations:

For Gemini, the parser handles markdown bold formatting (e.g., ****Female****) and full abstention phrases. For GPT-5.4 mini, whose outputs closely follow the specified format, a standard regex parser suffices. For Jais-2-8B-Chat, which exhibited higher format variability, we implement a progressive multi-pattern parser (v3) that handles six distinct output formats observed during experimentation, including reversed Reasoning/Answer order, parenthetical explanations, answer-first format, square-bracket abstentions, and Arabic label variants (*thakar/untha*) for the Arabic prompt condition.

Responses that cannot be parsed by any pattern are recorded as missing. Responses that explicitly state the gender cannot be determined are recorded as abstentions. Both missing and abstention cases are excluded from accuracy and F1 calculations and reported separately.

3.8 Evaluation Metrics

We report the following metrics for each experimental condition:

Accuracy: proportion of valid predictions matching the ground-truth gender label; this is the official ranking metric of the PAN 2018 shared task [1].

Macro F1: unweighted average of female and male F1 scores, treating both classes equally regardless of support and therefore sensitive to gender-asymmetric errors.

Abstention rate: proportion of authors/tweets for which the model explicitly declined to predict.

Missing rate: proportion of authors/tweets for which no parseable prediction could be extracted.

Per-gender recall: female and male recall reported separately to quantify gender bias in model predictions.

Recall gap: absolute difference between female and male recall, used as a summary measure of gender bias magnitude.

Statistical significance between model pairs is assessed using McNemar’s test [42] on common test examples, with continuity correction applied. We report chi-squared statistics and p -values, with significance thresholds at $p < 0.05$ (*), $p < 0.01$ (**), and $p < 0.001$ (***)

3.9 Reproducibility

All experiments use fixed random seeds (seed=42) for tweet sampling and batch assignment. Tweet-to-batch assignments are generated once and stored, ensuring identical experimental conditions across all models. API calls are logged with full raw responses for audit and re-analysis. Incremental saving with resume capability is implemented for all long-running experiments to ensure recoverability from interruptions. All code is available upon request.

4 Results

4.1 Author-Level Gender Classification

Table 1 presents the performance of all three LLMs in a zero-shot author-level gender classification setting on the PAN 2018 Arabic dataset, alongside the official competition baseline.

All three LLMs fall below both classical SVM baselines. Gemini 2.5 Flash and GPT-5.4 mini perform comparably at 78.54% and 78.20% respectively (a difference of only 0.34 points), while both fall 3.2–3.5 points below the best Arabic system (81.70%). Jais-2-8B-Chat performs substantially lower at 67.86%, a gap of 13.84 points from the best baseline and 10.68 points below Gemini. These results suggest that zero-shot LLMs do not yet surpass traditional feature-based methods on Arabic gender classification, a finding that contrasts sharply with results reported for English, where GPT-4 achieved 90.2% on the same benchmark family [5], compared to the English baseline of 82.21%. The English-Arabic capability gap is substantial: while LLMs exceed classical methods by approximately 8 points in English, they fall below them by approximately 3–3.5 points in Arabic.

Table 1 Author-Level Results vs. PAN 2018 Baseline

System	Approach	Valid	Accuracy	95% CI	Macro F1	Abstentions
Tellez et al. [17]	SVM + n-/skip-grams	1000/1000	81.70%	—	—	0%
Daneshvar & Inkpen [18]	SVM + n-grams	1000/1000	80.90%	—	—	0%
Gemini 2.5 Flash (ours)	Zero-shot LLM	983/1000	78.54%	[76.0, 81.0]	78.38%	1.5%
GPT-5.4 mini (ours)	Zero-shot LLM	1000/1000	78.20%	[75.8, 80.8]	77.60%	0%
Jais-2-8B-Chat (ours)	Zero-shot LLM	974/1000	67.86%	[64.9, 70.7]	67.52%	0.3%

95% confidence intervals computed via bootstrapping (1,000 iterations) on valid predictions only.

Bootstrapped 95% confidence intervals confirm these findings. Gemini [76.0%, 81.0%] and GPT-5.4 mini [75.8%, 80.8%] show substantially overlapping intervals, consistent with their statistically indistinguishable performance ($p = 1.00$). The confidence interval for Jais [64.9%, 70.7%] does not overlap with those of either multilingual model, providing additional confirmation that its lower performance represents a reliable difference rather than sampling variation.

The consistent underperformance of Jais-2-8B-Chat relative to both multilingual models is counterintuitive. Despite being trained primarily on Arabic data, Jais is outperformed by Gemini 2.5 Flash by 10.68 percentage points across all conditions. This suggests that **Arabic-native LLM training does not guarantee superior performance on Arabic NLP tasks**, and that the scale and diversity of multilingual training data may be more important than language-specific specialisation for this task. We note, however, that this comparison confounds training focus with model scale: Jais-2-8B-Chat is an 8-billion-parameter open-weight model, whereas the two commercial systems are presumed to be substantially larger, so this conclusion should be read as applying to Arabic-native models at the scale evaluated here.

4.2 Gender-Specific Performance and Bias at Author-Level

Table 2 presents the per-gender breakdown of classification performance at author level.

Table 2 Per-Gender Performance — Author-Level

Model	F. Prec.	F. Recall	M. Prec.	M. Recall	Gap	Bias
Gemini 2.5 Flash	0.75	0.86	0.84	0.70	16 pts	Female-leaning
GPT-5.4 mini	0.71	0.95	0.92	0.62	33 pts	Strong female-leaning
Jais-2-8B-Chat	0.65	0.78	0.73	0.57	21 pts	Female-leaning

A consistent **female-leaning bias** is observed across all three models at author level. GPT-5.4 mini shows the strongest bias (female recall 95%, male recall 62%, 33-point gap), followed by Jais (female recall 78%, male recall 57%, 21-point gap) and Gemini (female recall 86%, male recall 70%, 16-point gap). The consistency of

this bias across three models from different developers strongly suggests it is a **systemic property of LLMs processing Arabic text**, with significant implications for fairness in Arabic NLP applications.

4.3 Abstention Behaviour at Author-Level

The three models exhibit markedly different abstention behaviours at author-level. GPT-5.4 mini produced zero abstentions across all 1,000 authors. Gemini 2.5 Flash abstained on 15 authors (1.5%), while Jais-2-8B-Chat abstained on 3 (0.3%). Qualitative inspection of Gemini’s 15 abstentions revealed a consistent pattern: tweets consisted predominantly of gender-neutral religious content or formal political discourse. Abstentions were distributed asymmetrically: 11 of 15 (73.3%) occurred on male-authored profiles, suggesting female Arabic writing contains more gender-distinctive stylistic markers.

4.4 Tweet-Level Gender Classification

To examine LLM gender classification under minimal text signal, all three models were evaluated on individual tweets from the same 1,000 authors, under the fixed batching protocol described in Section 3.3 and both prompt variants (standard and forced binary).

Table 3 Tweet-Level Results — Standard vs. Forced Binary Prompt

Model	Prompt	Valid	Abstention%	Missing%	Accuracy	95% CI	Macro F1
Gemini 2.5 Flash	Standard	292/1000	70.8%	0.0%	73.29%	—	72.78%
Gemini 2.5 Flash	Forced Binary	1000/1000	0.0%	0.0%	64.50%	[61.5, 67.3]	64.50%
GPT-5.4 mini	Standard	161/1000	83.9%	0.0%	70.81%	—	64.47%
GPT-5.4 mini	Forced Binary	1000/1000	0.0%	0.0%	62.10%	[59.0, 65.0]	62.10%
Jais-2-8B-Chat	Standard	451/1000	47.3%	7.6%	60.53%	—	60.32%
Jais-2-8B-Chat	Forced Binary	876/1000	2.2%	10.2%	54.22%	[50.8, 57.3]	54.00%

95% CIs via bootstrapping (1,000 iterations) reported for forced binary only; standard prompt CIs omitted due to high abstention rates reducing effective sample size.

4.5 The Text-Volume Threshold Effect

Table 4 compares performance across text granularity levels for all three models.

A clear and consistent performance degradation is observed as available text decreases across all models. Gemini drops 5.25 points from author-level (78.54%) to tweet-level standard (73.29%), falling to 64.50% under forced binary. GPT drops from 78.20% to 62.10%, and Jais from 67.86% to 54.22%, approaching chance performance. This text-volume threshold effect is consistent across all three models, confirming it is a fundamental property of Arabic gender classification rather than a model-specific artefact. The only prior LLM data point on PAN 2018 Arabic gender, a preliminary multimodal GPT-4V probe on single image–text pairs (70.0%) [13], sits between our

Table 4 Performance vs. Available Text Signal — Complete Results

Model/Condition	Text	Valid	Accuracy	Macro F1	Abstention%
PAN 2018 best Arabic (SVM)	100 tweets	1000/1000	81.70%	—	0%
Gemini — Author-level	100 tweets	983/1000	78.54%	78.38%	1.5%
Gemini — Tweet standard	1 tweet	292/1000	73.29%	72.78%	70.8%
Gemini — Tweet forced	1 tweet	1000/1000	64.50%	64.50%	0%
GPT-5.4 mini — Author	100 tweets	1000/1000	78.20%	77.60%	0%
GPT-5.4 mini — Tweet std	1 tweet	161/1000	70.81%	64.47%	83.9%
GPT-5.4 mini — Tweet forced	1 tweet	1000/1000	62.10%	62.10%	0%
Jais — Author-level	100 tweets	974/1000	67.86%	67.52%	0.3%
Jais — Tweet standard	1 tweet	451/1000	60.53%	60.32%	47.3%
Jais — Tweet forced	1 tweet	876/1000	54.22%	54.00%	2.2%

author-level and tweet-level results, directionally consistent with this account, though modality and sampling differences preclude direct comparison.

GPT-5.4 mini’s abstention rate changes most sharply: it abstains on 0% of authors at author level but **83.9%** of tweets at tweet level, the largest abstention jump of any model (+83.9 points), compared to +69.3 points for Gemini and +47.0 points for Jais. By this measure, GPT is the most conservatively calibrated of the three models.

4.6 Abstention Behaviour and Model Calibration

To validate whether tweet-level abstentions reflect genuine uncertainty, we examined model accuracy, under forced prediction, on the 708 tweets on which Gemini had previously abstained. Accuracy on this subset collapsed to 60.88%, significantly lower than the 73.29% achieved on voluntarily committed tweets, confirming that abstentions represent accurate self-assessment of classification difficulty rather than over-caution.

Abstentions at tweet level were distributed in a near-gender-balanced manner across all models (Gemini: male 51.6%, female 48.4%; GPT: male 53.2%, female 46.8%), in contrast to the author-level abstention pattern where 73.3% of Gemini’s abstentions fell on male-authored profiles. This confirms that tweet level abstentions are driven by general text sparsity rather than gender-specific ambiguity.

Tweet length analysis revealed model-specific patterns. For Gemini and Jais, committed tweets were measurably longer than abstained tweets (Gemini: 78.6 vs. 70.6 characters, 11.3% difference; Jais: 74.5 vs. 70.1 characters, 6.2% difference). In contrast, GPT showed no systematic length difference (−2.2 characters, −3.0%), suggesting GPT’s abstention mechanism is driven primarily by content type rather than text volume. Among Gemini’s committed tweet-level predictions, accuracy was highest for the shortest tweets (0–50 chars: 77.6%) and decreased with length (51–100 chars: 73.6%; 101–150 chars: 69.1%).

4.6.1 Religious Content Analysis

The calibration finding above raises a further question: what specific content types drive tweet-level abstentions? To address this, we conducted a three-level analysis of

religious content, the largest single abstention category identified through qualitative inspection, directly validating whether models abstain on genuinely gender-ambiguous content or do so indiscriminately. The analysis uses three levels: (1) presence of terms from a curated 147-term Islamic religious lexicon across seven semantic categories (divine references, supplications, scripture, prayer, pillars of Islam, jurisprudence, and afterlife vocabulary), provided in full in Appendix A; (2) token overlap with a reference corpus of 6,236 Quranic verses [43] and 50,637 authenticated Hadiths, obtained from the 50,762-record, 14-book collection of [44] after excluding entries with missing Arabic text; and (3) religious token density within each tweet, defined as the proportion of tweet tokens matching the religious lexicon or reference corpus. All three levels converge on the same conclusion: it is not the presence of religious vocabulary per se but its dominance of tweet content that triggers abstention, directly confirming that models abstain on content that genuinely lacks gender-discriminative signal, consistent with the calibration finding above.

Presence level. At the level of religious vocabulary presence, religious terms appeared in 81–88% of both abstained and committed tweets, with negligible enrichment ratios ($0.97\text{--}1.06\times$). This finding reveals that Arabic religious vocabulary functions as a pervasive discourse marker in Arabic social media rather than a reliable abstention predictor: the mere presence of religious terms in a tweet is insufficient to trigger abstention, as such terms appear in the vast majority of Arabic tweets regardless of topic.

Density level. A substantially stronger pattern emerged at the density level. Table 5 presents the results of the density analysis across all three models.

Table 5 Religious Content Density Analysis — Abstained vs. Committed Tweets

Model	Avg. Density		Density Ratio	High-Density Rate ($>40\%$)	
	Abstained	Committed		Abstained	Committed
Gemini 2.5 Flash	0.373	0.276	$1.35\times$	46.19%	23.97%
GPT-5.4 mini	0.355	0.296	$1.20\times$	41.72%	30.43%
Jais-2-8B-Chat	0.367	0.319	$1.15\times$	43.97%	35.48%

Enrichment ratios for high-density rate: Gemini $1.93\times$, GPT $1.37\times$, Jais $1.24\times$.

Tweets where religious terms constituted more than 40% of total tokens were $1.24\text{--}1.93\times$ more likely to trigger abstention than commitment, with Gemini showing the strongest effect ($1.93\times$). This confirms that the abstention trigger is not religious vocabulary per se, but **predominantly religious text**: tweets consisting mainly of Quranic supplications, dhikr, or Islamic invocations, which provides no gender-discriminative stylistic signal regardless of author gender.

Term-level analysis. The term *Allah* appeared in 22.9–24.9% of abstained tweets, compared to 10.6–18.6% of committed tweets, a token-frequency enrichment ratio of $1.3\text{--}2.3\times$. This token-frequency measure differs from the presence-level analysis above, where religious vocabulary appeared in 81–88% of both abstained and committed tweets, reflecting the pervasiveness of religious vocabulary in Arabic Twitter generally.

The other most frequent religious terms in abstained tweets across all models were *Allahumma* (8–9%), *subhan* (2%), *rabbi* (2%), and *du’a* (2%), confirming that the abstention-triggering content is primarily supplication and devotional language rather than religious discussion or commentary.

4.6.2 Non-Religious Abstention Profiles

Among the 53.8% of Gemini’s abstained tweets classified as non-predominantly-religious (religious token density <40%), qualitative inspection revealed three secondary abstention profiles:

Generic social exchanges. Simple thank-you messages, isolated hashtags (e.g., *#FridayBlessing*), and brief conversational exchanges provide insufficient content for gender inference regardless of religious content.

Dialectal Arabic. Tweets in Libyan, Sudanese, Gulf, and Egyptian dialects exhibit non-standard morphology and lexicon that reduce the availability of gender-distinctive cues for models primarily trained on Modern Standard Arabic [45].

Formal institutional discourse. Political commentary, news sharing, and civic announcements use an impersonal register that provides no personal stylistic signal for gender inference.

Together, these findings characterise Arabic tweet-level abstentions as driven by four distinct content types: (1) high-density religious text, (2) generic social exchanges, (3) dialectal Arabic, and (4) formal impersonal discourse, all sharing the property of providing insufficient gender-discriminative signal for reliable classification.

4.7 Gender Bias Across All Conditions

Table 6 presents the complete per-gender recall across all models and experimental conditions.

Table 6 Female vs. Male Recall Across All Models and Conditions

Model	Condition	F. Recall	M. Recall	Gap	Bias
Gemini 2.5 Flash	Author-level	0.86	0.70	16 pts	Female-leaning
Gemini 2.5 Flash	Tweet standard	0.79	0.66	13 pts	Female-leaning
Gemini 2.5 Flash	Tweet forced	0.65	0.64	1 pt	Negligible
GPT-5.4 mini	Author-level	0.95	0.62	33 pts	Strong female-leaning
GPT-5.4 mini	Tweet standard	0.85	0.43	42 pts	Strong female-leaning
GPT-5.4 mini	Tweet forced	0.62	0.62	0 pts	None
Jais (EN prompt)	Author-level	0.78	0.57	21 pts	Female-leaning
Jais (EN prompt)	Tweet standard	0.64	0.57	7 pts	Female-leaning
Jais (EN prompt)	Tweet forced	0.62	0.46	16 pts	Female-leaning
Jais (AR prompt)	Tweet standard	0.63	0.68	5 pts	Male-leaning
Jais (AR prompt)	Tweet forced	0.37	0.75	38 pts	Strong male-leaning

Three key patterns emerge. **First, female-leaning bias is universal under English prompting:** every English-prompt condition for every model shows higher female recall, without exception. **Second, forced binary prompting eliminates the female-leaning bias for Gemini and GPT:** Gemini’s gap collapses from 16 to 1 point, GPT’s from 33 to 0 points, but Jais retains a 16-point gap even under forced conditions. **Third, Arabic prompts reverse the bias direction for Jais:** the Arabic forced prompt produces the most extreme bias of any condition: a 38-point male-leaning gap (female recall 37%, male recall 75%), directly opposite to the female-leaning bias under English prompting.

The progressive attenuation of female-leaning bias as text volume decreases further supports the hypothesis that the bias is confidence-driven: as available evidence decreases, models become less confident in female-leaning associations, producing more balanced predictions. Deployment contexts that force binary predictions may paradoxically produce more gender-fair outcomes than those allowing models to express confidence.

4.8 Prompt Language Study: English vs. Arabic Prompts on Jais-2-8B-Chat

4.8.1 Motivation

As reviewed in Section 2.4, existing prompt-language comparisons, including those covering Arabic-centric models [6], report aggregate accuracy only; how prompt language shapes an Arabic-native model’s task behaviour (coverage, calibration, format compliance, and bias direction) remains unexamined. This section reports the first such behavioural analysis for an Arabic-native LLM.

4.8.2 Experimental Setup

Jais-2-8B-Chat was examined under English and Arabic prompt conditions using the same fixed tweet-to-batch assignment (seed=42). The system prompt, task description, format instructions, and output labels were all translated to Arabic in the Arabic condition. Output labels were translated as: Male→*thakar* (masculine), Female→*untha* (feminine), Cannot be determined→*la yumkin tahdiduh*. Both standard and forced binary variants were examined under each language condition, yielding four experimental conditions.

4.8.3 Results

Table 7 presents the results of the prompt language comparison across all four conditions. Arabic standard prompting produced the highest per-prediction accuracy (65.22%) but the lowest coverage (18.4% valid predictions), while Arabic forced prompting produced the lowest accuracy (55.77%) and highest format non-compliance (16.8% missing). The abstention rate under Arabic standard prompting (81.2%) was 33.9 points higher than under English standard prompting (47.3%).

Table 7 Prompt Language Comparison — Jais-2-8B-Chat (Tweet-Level)

Condition	Valid	Abstention%	Missing%	Accuracy	Macro F1	Bias
English — Standard	451/1000	47.3%	7.6%	60.53%	60.32%	Female-leaning
English — Forced binary	876/1000	2.2%	10.2%	54.22%	54.00%	Female-leaning
Arabic — Standard	184/1000	81.2%	0.4%	65.22%	64.88%	Male-leaning
Arabic — Forced binary	832/1000	0.0%	16.8%	55.77%	54.16%	Strong male-leaning

4.8.4 Discussion

Accuracy vs. coverage trade-off. Arabic prompts produced higher accuracy on committed predictions (+4.69 points: 65.22% vs. 60.53%) at the cost of sharply reduced coverage: only 18.4% of tweets received valid predictions vs. 45.1% under English prompting. The abstention rate increased from 47.3% to 81.2%.

Minimal effect under forced binary conditions. When abstentions were disallowed, prompt language had negligible effect on accuracy (54.22% vs. 55.77%, $p = 0.227$, not significant). The standard-prompt accuracy advantage is driven by selective abstention rather than genuine capability improvement.

Bias reversal under Arabic prompting. Arabic prompts reverse Jais’s bias direction: while English prompts produce female-leaning predictions, Arabic prompts consistently produce male-leaning predictions. The Arabic forced prompt exhibits the most extreme bias of any tested condition: female recall 37% vs. male recall 75%, a 38-point gap. This prompt-language-dependent bias reversal suggests Jais’s internal gender associations are fundamentally sensitive to the language of instruction.

Format compliance. Arabic forced prompts produced higher format non-compliance (16.8% missing vs. 10.2%), despite Jais being Arabic-native, suggesting that structured output compliance is more strongly developed in the English portion of Jais’s training data. This interpretation is independently corroborated by the model developers’ own instruction-following evaluation: on IFEval, Jais-2-8B scores 63.14 under the English-language benchmark variant but only 58.17 under the Arabic-language variant (prompt-level strict accuracy) [29], indicating that even an Arabic-centric model complies with structured instructions more reliably in English.

4.8.5 Implications

These findings challenge the narrative that Arabic-native models perform better when prompted in Arabic. Arabic prompts improve precision but reduce coverage and produce severe bias reversals. Neither English nor Arabic prompts are bias-free: English produces female-leaning bias, Arabic produces male-leaning bias.

4.9 Inter-Model Agreement and Statistical Significance

4.9.1 Inter-Model Error Agreement

To assess whether models share failure cases or fail independently, we analysed prediction agreement across the 960 authors for which all three models produced valid predictions.

Table 8 Inter-Model Error Agreement
(Author-Level, $n = 960$)

Error Pattern	Count	%
All 3 models correct	490	51.0%
All 3 models wrong	64	6.7%
Only Gemini wrong	78	8.1%
Only GPT-5.4 mini wrong	50	5.2%
Only Jais wrong	155	16.1%

Only 51.0% of the dataset is reliably classified by all three models. Jais makes by far the most unique errors: 155 authors (16.1%) that both Gemini and GPT classify correctly, compared to 78 for Gemini and 50 for GPT, confirming that Jais’s errors represent a qualitatively different error profile.

Pairwise prediction agreement was highest between Gemini and GPT (77.4%), followed by GPT and Jais (71.9%), and lowest between Gemini and Jais (65.9%), consistent with the two multilingual models sharing similar internal representations of Arabic gender markers.

Of the 64 authors misclassified by all three models unanimously, **59 (92.2%) were male-authored profiles that all three models predicted as female**, with only 5 female-authored profiles receiving unanimous incorrect male predictions. This near-perfect cross-model agreement on the direction of error confirms that certain male Arabic writing styles are systematically unrecognisable as male to current LLMs.

4.9.2 Statistical Significance Testing

McNemar’s test [42] was applied to all model pairs on common test examples to determine whether accuracy differences are statistically reliable.

Gemini and GPT-5.4 mini are statistically indistinguishable at both author-level ($p = 1.00$) and tweet-level ($p = 0.087$), constituting a single multilingual LLM performance tier with a ceiling of approximately 78–79% on Arabic zero-shot gender classification.

Jais is significantly worse than both multilingual models at all levels ($p < 0.001$ in all four comparisons). The performance gap is statistically robust.

The Arabic prompt accuracy advantage for Jais is not statistically significant ($p = 0.424$ standard, $p = 0.227$ forced). The observed accuracy difference is better attributed to selective abstention behaviour than genuine capability improvement.

Table 9 McNemar’s Test Results

Comparison	N	χ^2	p -value	Sig.
Gemini vs GPT-5.4 mini (author)	983	0.00	1.000	ns
Gemini vs Jais (author)	957	30.25	<0.001	***
GPT-5.4 mini vs Jais (author)	974	36.36	<0.001	***
Gemini vs GPT (tweet forced)	1000	2.94	0.087	ns
Gemini vs Jais (tweet forced)	876	24.62	<0.001	***
GPT vs Jais (tweet forced)	876	12.48	<0.001	***
Jais EN vs AR standard	129	0.64	0.424	ns
Jais EN vs AR forced	732	1.46	0.227	ns

*** $p < 0.001$; ns = not significant

4.10 Qualitative Error Analysis and Bias Mechanisms

4.10.1 Stereotypical Reasoning in Misclassifications

To understand the mechanisms driving the female-leaning bias, we conducted a keyword analysis of model reasoning in misclassified cases. Among GPT-5.4 mini’s 191 male-authored profiles incorrectly classified as female, the term “feminine” appeared in 72.8% of reasoning texts, “expressive” in 62.8%, and “emotional” in 59.2%. In nearly 3 out of 4 misclassification cases, the model explicitly cited feminine characteristics, indicating it has internalised gender stereotypes as classification heuristics rather than relying on direct linguistic evidence.

4.10.2 Three Distinct Misclassification Profiles

Religious content ambiguity. A substantial subset of misclassified male authors wrote predominantly religious content: Quranic verses, Islamic supplications, and dhikr. This content is stylistically formulaic with no gender-distinctive markers, causing models to default to female predictions. This is consistent with the abstention analysis (Section 4.6), where high-density religious tweets were $1.24\text{--}1.93\times$ more likely to trigger abstention than commitment across all models.

Arabic morphological confusion. Several misclassified male authors wrote tweets addressing female interlocutors using Arabic feminine verb conjugations. Models systematically interpreted these *addressee-gender* markers as *author-gender* markers, a fundamental Arabic morphological error. A male author writing to a female friend uses feminine address forms that reflect the grammatical gender of the addressee, not the author [45]. This error type represents a specific, linguistically grounded weakness in LLM handling of Arabic grammatical gender morphology that has no direct English analogue.

Stereotypical topic and style associations. The most prevalent misclassification profile involves topic-based stereotyping. Male authors who expressed admiration for female celebrities, shared sentimental quotes, used warm or supportive communication styles, or posted with emotional emojis were consistently misclassified as female.

GPT’s reasoning explicitly cited the “feminine” quality of these topics, revealing internalised cultural stereotypes about gender-appropriate topics and tones that do not reliably hold in practice.

The three profiles share a common structure: in each case, the model has *correctly identified* the relevant linguistic feature (formulaic religious content, feminine verb forms, emotional fan language), but has *incorrectly attributed* that feature to the author’s gender rather than to the content type, the addressee’s gender, or a culturally widespread male behaviour. The misclassification mechanism is therefore not a failure of linguistic observation but a failure of inference: the model observes the correct signal but applies an incorrect gender association to it.

4.10.3 Reasoning Excerpts Across Misclassification Profiles

Table 10 presents representative reasoning excerpts from GPT-5.4 mini for one misclassified male author from each profile. These excerpts illustrate the specific linguistic features each model cites as evidence of female authorship, and the specific reasoning failure each represents.

In the third profile, GPT-5.4 mini explicitly states that “men can certainly write this way” before nonetheless predicting female: the model’s probabilistic gender priors override its own reasoning. We analyse this *stereotype override* pattern and its implications for debiasing in Section 5.3.

4.10.4 Implications for Arabic NLP

These three failure modes have distinct remediation implications. Religious content ambiguity is addressable with more text, as demonstrated by the author-level performance improvement over tweet-level. The morphological confusion finding suggests that explicit Arabic morphological preprocessing to disambiguate addressee-gender from author-gender markers, a task for which mature Arabic NLP tools exist [45], could substantially reduce this error type without retraining. The stereotypical association finding, however, reflects biases deeply embedded in pretraining data that are unlikely to be resolved without targeted debiasing interventions specifically designed for Arabic’s cultural and linguistic context.

5 Discussion

This study set out to answer four research questions concerning the capabilities, reliability, fairness, and Arabic-specific behaviour of LLMs on author gender profiling. Our empirical findings provide clear, grounded answers to all four, and reveal several previously undocumented phenomena along the way.

5.1 The English-Arabic LLM Capability Gap

Our most fundamental finding is that zero-shot LLMs do not surpass the classical SVM baseline on Arabic gender classification, despite doing so dramatically in English. Gemini 2.5 Flash and GPT-5.4 mini both fall 3.2–3.5 points below the best PAN 2018 Arabic system (81.70%), while Jais-2-8B-Chat falls 13.84 points below it. This

Table 10 Representative GPT-5.4 mini Reasoning Excerpts for the Three Misclassification Profiles (all authors are male, all predicted female)

Profile	Model Reasoning (excerpt)	Reasoning Failure
Religious content ambiguity	“The tweets are overwhelmingly religious dhikr, supplications, and copied prophetic invocations. That makes gender inference difficult because the content is highly formulaic and not strongly gendered. Still, there are a few weak cues: the account repeatedly posts short devotional phrases in a very consistent, almost automated style.”	The model correctly identifies the content as formulaic and gender-neutral, yet still defaults to a female prediction. Formulaic religious content provides no discriminative signal, yet the model is biased toward female when uncertain.
Arabic morphological confusion	“There are several indicators pointing strongly toward female: the author directs much speech in the feminine form to others, such as <i>tak-tubi</i> , <i>tazawwajihi</i> , <i>tarsumihi</i> — this suggests the speaker is accustomed to addressing women or is a woman within a female social context.”	The model interprets feminine Arabic verb conjugations (addressee-gender markers) as evidence of the author’s own gender. A male author addressing a female interlocutor uses feminine forms that reflect the <i>addressee’s</i> gender, not his own, a fundamental Arabic morphological error.
Stereotypical associations	“The biggest clue is the intense fan-style admiration for the female singer Elissa, using affectionate language like <i>habibi</i> , <i>malkat al-ihsas</i> , <i>najmtuna</i> , and repeated emotional support with hearts and kiss emojis. While men can certainly write this way, the overall tone here is highly expressive, emotionally supportive, and socially relational.”	The model explicitly acknowledges men can write this way, yet still predicts female, demonstrating that probabilistic gender priors override explicit reasoning. This represents a form of stereotype override [10] in which cultural associations dominate over direct linguistic evidence.

contrasts sharply with the English result reported by Staab et al. [5], where GPT-4 exceeded the English baseline by approximately 8 points. The combined gap between English and Arabic LLM performance relative to their respective baselines is therefore approximately 11 points, representing a substantial English-Arabic capability deficit.

Several factors likely contribute to this gap. Arabic’s morphological complexity (grammatical gender marking, dialectal variation, and code-switching with English) creates classification signals that are more subtle and context-dependent than in English [45]. Furthermore, Arabic is underrepresented in the pretraining data of multilingual models relative to English, resulting in weaker stylistic representations for Arabic text [22, 46]. Arabic-specific pretrained models such as MARBERT [22], trained

on 1 billion Arabic tweets, demonstrate the potential of large-scale Arabic-specific pretraining for social media NLP tasks, but their discriminative training paradigm is fundamentally different from the generative zero-shot setting examined here, suggesting the gap may be specific to zero-shot generation rather than Arabic language understanding more broadly. Notably, however, reported results for *fine-tuned* Arabic transformers on related PAN author-profiling splits (77–79.9% [20]) also sit at or below the official shared-task SVM results, hinting that the difficulty of this task may not be specific to the zero-shot setting — though differences in evaluation protocol preclude a direct comparison. Our finding that the Arabic-native Jais model performs substantially worse than both multilingual models further challenges the intuition that Arabic-first training resolves these limitations: it appears that the scale and diversity of multilingual pretraining data outweigh the benefit of Arabic-focused specialisation for this task, at least at the 8-billion-parameter scale tested here.

These results have important practical implications. Systems that deploy LLMs for Arabic author profiling in the zero-shot setting cannot rely on the performance levels achievable in English, and should consider fine-tuning or few-shot adaptation as necessary complements to zero-shot prompting.

5.2 Calibrated Abstention as a Reliability Signal

Our calibration analysis reveals that LLM abstentions are genuinely informative rather than excessively conservative. When Gemini’s previously-abstained tweets were forced to classification, accuracy collapsed to 60.88% (barely above chance on a binary task), confirming that the model correctly identified cases it could not reliably classify. This calibration property is consistent across models, with GPT-5.4 mini showing the most precise calibration: it abstains on 0% of author-level cases (where 100 tweets provide ample signal) but 83.9% of tweet-level cases (where a single tweet often provides insufficient signal).

Religious content emerges as a universal abstention trigger across all models. The term *Allah* appeared as a proportion of tweet tokens $2.2\text{--}2.3\times$ more frequently in abstained than committed tweets under keyword frequency analysis; however, this reflects token-level frequency rather than tweet-level presence. As shown in Table 5, when measured at the tweet presence level, religious vocabulary appeared in 81–88% of both abstained and committed tweets (enrichment ratio $0.97\text{--}1.06\times$), while high-density religious tweets ($>40\%$ religious tokens) showed a stronger abstention enrichment of $1.24\text{--}1.93\times$ across models. Together these three measurements converge on the same conclusion: it is the *density* of religious content, not its mere presence, that distinguishes abstained from committed tweets. This finding reflects a genuine linguistic challenge: Islamic religious text is highly formulaic, gender-neutral, and stylistically uniform across authors, providing no discriminative signal for gender inference. Importantly, this is not a model failure but an appropriate uncertainty response to genuinely ambiguous content.

5.3 Systemic Female-Leaning Bias and Its Modulation

A consistent female-leaning bias emerges across all three models under English prompting: female authors are recalled more reliably than male authors, with gaps ranging from 7 to 42 percentage points depending on model and experimental condition. The fact that this bias appears across three models from different developers, with different architectures and training data, strongly suggests it is a systemic property of LLMs processing Arabic text rather than an artefact of any individual model.

Two mechanisms appear to drive this bias. First, qualitative analysis of misclassification reasoning reveals that LLMs explicitly associate emotional expressiveness, warm tone, and fan-like admiration with female authorship, applying cultural gender stereotypes absorbed from pretraining data. Second, Arabic morphological structure creates a unique failure mode: male authors writing to female interlocutors use feminine grammatical forms that LLMs systematically misread as signals of the author’s own gender. This morphological confusion finding is linguistically specific to Arabic and has no direct English analogue, highlighting the need for Arabic-specific bias analysis that goes beyond adapting English-language frameworks.

The modulation of bias by experimental condition is equally revealing. The female-leaning bias diminishes as text volume decreases, and disappears entirely under forced binary prompting for Gemini and GPT-5.4 mini. This pattern suggests that gender bias in these models is not a fixed classification property but is linked to the models’ confidence mechanisms: when models commit voluntarily, they draw on stereotypical stylistic associations; when forced to predict without confidence, these associations dissipate and predictions become near-random. Paradoxically, deployment contexts that eliminate abstention options, such as automated classification pipelines, may produce more gender-fair outcomes than those that allow models to express confidence. The reasoning excerpts in Table 10 (Section 4.10.3) provide direct evidence for this mechanism at the individual prediction level: the model can acknowledge that men write this way and still predict female, culturally absorbed associations overriding direct linguistic evidence even when the model’s own analysis contradicts them. This *stereotype override* has been documented in English LLMs [9, 10] but is here observed for the first time in Arabic, suggesting that it is a cross-lingual property of LLM gender reasoning rather than an English-specific artefact. Furthermore, the morphological confusion profile reveals a second, distinctly Arabic mechanism: models correctly identify feminine verb forms in the text but misattribute them to the author rather than the addressee, an error that requires Arabic morphological awareness to avoid, and that debiasing approaches developed for English cannot address [45].

5.4 Arabic Prompts and Bias Reversal in Jais

Our prompt language study reveals that the English-prompt advantage documented for multilingual LLMs [6, 7, 12] does not translate straightforwardly to an Arabic-native model. For Jais-2-8B-Chat, Arabic prompts improve accuracy on committed predictions (+4.69 points) at the cost of sharply reduced coverage (abstention rate 81.2% vs. 47.3% under English), and the accuracy advantage is not statistically significant ($p = 0.227$). Beyond accuracy, Arabic prompts reverse the bias direction: from

female-leaning under English prompts to male-leaning under Arabic prompts, with the Arabic forced condition producing the most extreme bias of any tested setting (38-point male-leaning gap).

This bias reversal suggests that Jais’s internal gender representations are sensitive to prompt language in a way that goes beyond simple performance differences. One possible interpretation is that the model’s Arabic-language representations are more strongly tied to traditional cultural gender associations in Arabic discourse, while its English-language representations reflect the more gender-equal implicit associations in English-language pretraining data. Whatever the mechanism, the practical implication is clear: neither English nor Arabic prompts are bias-free for Jais, and practitioners must choose between female-leaning (English) and male-leaning (Arabic) bias depending on their application’s fairness requirements.

5.5 Limitations

Several limitations of this study should be acknowledged. First, our study is restricted to a single dataset (PAN 2018 Arabic), which consists exclusively of Twitter content from a specific time period (2018). Generalisation to other Arabic text genres (news, literary text, conversational speech) or to more recent social media data cannot be assumed. The Twitter-specific nature of the dataset means that our findings on religious content abstention, dialectal variation, and gender bias patterns may reflect the particular register and demographics of Arabic Twitter users in 2018 rather than Arabic language use more broadly. Future work should evaluate on PAN 2017 Arabic, the gender-labelled MSA news article dataset of Alshaalan and Al-Khalifa [23], and newly-collected dialectal Arabic corpora to assess generalisation of our findings across datasets, time periods, and text genres. Second, we study only three LLMs; while these span a range of architectures and training paradigms, the field is evolving rapidly and results may differ for other models. In particular, the Arabic-native model we evaluate is an 8-billion-parameter open-weight model, substantially smaller than the two commercial systems, so the observed Jais deficit conflates language specialisation with model scale; we note that pairing open-weight models in the 7–9B range with proprietary frontier systems as a state-of-the-art reference mirrors the design of recent Arabic LLM evaluations [27]. Evaluating the larger Jais-2-70B-Chat would nonetheless disentangle these two factors and is left for future work. Third, our qualitative error analysis is based on a sample of misclassified cases and the keyword frequencies in model reasoning, which provides indicative rather than exhaustive characterisation of bias mechanisms. Fourth, we examine only zero-shot prompting; few-shot adaptation may yield substantially different performance and bias profiles. Fifth, the PAN 2018 dataset uses binary gender labels and does not account for non-binary gender identities, a limitation inherited from the original benchmark. Sixth, although multiple lines of evidence argue against memorisation-driven results (Section 3.1), the public availability of PAN 2018 means that indirect pretraining contamination cannot be fully excluded; relatedly, our batched tweet-level protocol, while shuffled across authors, cannot entirely rule out within-batch context effects.

5.6 Future Work

Several directions follow from this study. First, few-shot evaluation with carefully selected examples, particularly examples balanced across gender and dialect, could improve performance while potentially modulating bias in controllable ways. Second, the morphological confusion finding points to a concrete engineering intervention: integrating an Arabic morphological disambiguator [45] as a preprocessing step before LLM inference to explicitly mark addressee-gender versus author-gender verb forms; whether such preprocessing reduces this error type is an empirical question with direct practical implications. Third, the dialectal abstention profile identified in Section 4.6.2 — where tweets in Libyan, Sudanese, Gulf, and Egyptian dialects triggered disproportionate abstentions — suggests, in line with emerging dialectal evaluation resources [47], that dialect-aware analysis of Arabic author profiling is an important open problem. Fourth, a systematic qualitative coding study of LLM misclassification reasoning across a larger sample would provide a more complete taxonomy of bias mechanisms and could inform the design of debiasing interventions tailored to Arabic’s morphological gender system. Finally, the Arabic psychographic profiling task (inferring personality traits from Arabic text) represents a natural extension with limited prior LLM studies.

6 Conclusion

This paper presented, to the best of our knowledge, the first comprehensive empirical investigation of text-based large language model behaviour for Arabic author gender profiling on the official PAN 2018 test partition, addressing a critical gap that existing work had left open. We studied three state-of-the-art LLMs (Gemini 2.5 Flash, GPT-5.4 mini, and the Arabic-native Jais-2-8B-Chat) on the PAN 2018 Arabic author profiling benchmark under a multi-dimensional experimental framework crossing text granularity, prompt variant, and prompt language to examine performance, calibration, bias, and prompt sensitivity. Beyond classification performance, we conducted systematic bias analysis, abstention calibration testing, inter-model agreement analysis, statistical significance testing, and qualitative error characterisation, a depth of analysis not previously applied to Arabic author gender profiling.

Our results reveal five main findings. First, all three LLMs fall below the best classical SVM baseline of 81.70%, with Gemini and GPT performing comparably at approximately 78% and Jais substantially lower at 67.86%. This contrasts with the English setting where LLMs exceed classical methods by approximately 8 points, establishing a significant English-Arabic LLM capability gap. Second, a text-volume threshold effect governs abstention behaviour: all models show far higher abstention at tweet-level (47.3–83.9%) than at author level (0–1.5%), and forced predictions on previously-abstained content collapse to near-chance accuracy (60.88%), confirming that LLMs accurately self-assess classification difficulty in Arabic.

Third, a consistent female-leaning bias is observed across all three models under English prompting, partially explained by LLMs’ tendency to associate emotional expressiveness and specific topics with female authorship. Fourth, Arabic prompts for the Arabic-native Jais-2-8B-Chat improve prediction precision but reduce coverage

and reverse the bias direction from female-leaning to strongly male-leaning (38-point gap), the first documented prompt-language-dependent bias reversal in an Arabic LLM. Fifth, qualitative error analysis identifies three distinct misclassification profiles (religious content ambiguity, Arabic morphological confusion, and stereotypical topic associations), in which models correctly observe the relevant linguistic features but apply incorrect gender inferences. A form of *stereotype override*, in which gender priors overrule the model’s own explicit reasoning, is documented for the first time in Arabic; the morphological confusion profile has no English analogue and calls for Arabic-aware preprocessing rather than adapted English debiasing frameworks.

Together, these findings challenge two widely held assumptions: that LLMs generalise their English performance advantages to Arabic, and that Arabic-native LLMs outperform multilingual models on Arabic tasks. Both assumptions are contradicted by our results (the latter at the 8-billion-parameter scale evaluated here), with implications for the deployment of LLMs in Arabic NLP systems and for the design of Arabic-specific empirical research frameworks. A complementary contribution of this work is the release of a curated 147-term Arabic Islamic religious lexicon across seven semantic categories, which we make available as a supplementary resource for future Arabic NLP research on religious content detection.

All experimental code, prompt templates, parsed results, and analysis scripts are available upon request to support reproducibility and future work in this area.

Appendix A Islamic Religious Lexicon for Arabic NLP

Table A1 presents the curated 147-term Arabic Islamic religious lexicon developed for this study, organised by semantic category. Terms were selected based on their unambiguous religious specificity in Arabic social media discourse. Ambiguous terms that also carry secular meanings (e.g., *al-karim* [generous], *al-hakim* [wise]) were excluded from the core lexicon and handled separately via co-occurrence rules (Section 4.6.1). The lexicon is released as a supplementary resource to support future Arabic NLP research on religious content detection.

Statements and Declarations

- **Funding:** Not applicable.
- **Competing Interests:** The author has no relevant financial or non-financial interests to disclose.
- **Ethics approval:** Not applicable.
- **Consent for publication:** Not applicable.
- **Data availability:** The PAN 2018 dataset is publicly available through the PAN shared task repository at <https://pan.webis.de/clef18/pan18-web/author-profiling.html>. Experimental results and analysis scripts are available from the corresponding author upon reasonable request.
- **Code availability:** All experimental code, prompt templates, and parsing scripts are available from the corresponding author upon reasonable request.

Table A1 Curated Arabic Islamic Religious Lexicon (147 terms, 7 categories). Full Arabic lexicon available as supplementary material.

Category	Representative Terms (transliterated)
Divine references (16)	Allah, Allahumma, al-Rabb, Rabbi, Rabbuna, al-Rahman, al-Rahim, al-Qadir, al-Ghafur, al-Latif
Supplications & dhikr (23)	Alhamdulillah, Subhan, Bismillah, Astaghfirullah, Amin, Masha'Allah, In sha'Allah, Jazak, Barakallah, Ya Rabb, Subhanaka, La ilaha, Du'a, Dhikr
Scripture (15)	al-Quran, Aya, Sura, al-Fatiha, al-Baqara, al-Hadith, Hadith, Ahadith, al-Sunna, al-Sharifa, Tafsir
Prayer & worship (25)	al-Salah, Salah, al-Fajr, al-Zuhr, al-Asr, al-Maghrib, al-Isha, al-Jumu'a, al-Adhan, al-Imam, al-Masjid, Sujud, Wudu, al-Qibla
Pillars of Islam (13)	Zakat, al-Siyam, Sawm, Ramadan, al-Hajj, al-Umra, al-Islam, al-Iman, Iman, al-Shahada, al-Tawhid
Jurisprudence / fiqh (20)	Halal, Haram, Makruh, Mustahabb, Mubah, Wajib, Fard, Ja'iz, al-Fiqh, Fatwa, al-Shari'a, Muharram, al-Ta'a, al-Faqih
Prophet & companions (14)	al-Nabi, al-Rasul, Muhammad, Salla Allahu alayhi wa salam, Alayhi al-Salam, al-Sahaba, al-Khulafa, al-Rashidin, al-Mu'minin, al-Muslimin
Afterlife (14)	al-Akhira, al-Janna, al-Jahannam, al-Nar, al-Hisab, al-Qiyama, al-Ba'th, al-Sirat, al-Mizan, al-Tawba, al-Maghfira, Thawab, Iqab
Islamic calendar (7)	Sha'ban, Muharram, Dhu al-Hijja, Eid al-Fitr, Eid al-Adha, Laylat al-Qadr, al-Isra

- **Author contribution:** Single author. All work performed by T. Alrashid.
- **Declaration of generative AI use:** During the preparation of this work the author used an AI assistant (Claude, Anthropic) to assist with language editing, verification and correction of bibliographic references, and drafting revisions to portions of the manuscript text. After using this tool, the author reviewed, verified, and edited all content and takes full responsibility for the content of the published article.

References

- [1] Rangel, F., Rosso, P., Montes-y-Gómez, M., Potthast, M., Stein, B.: Overview of the 6th author profiling task at PAN 2018: Multimodal gender identification in Twitter. In: CLEF 2018 Labs and Workshops, Notebook Papers (2018)

- [2] Mikros, G.: Large language models and forensic linguistics: Navigating opportunities and threats in the age of generative AI. arXiv preprint arXiv:2512.06922 (2025)
- [3] Rangel, F., Rosso, P., Potthast, M., Stein, B.: Overview of the 5th author profiling task at PAN 2017: Gender and language variety identification in Twitter. In: CLEF 2017 Labs and Workshops, Notebook Papers (2017)
- [4] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in Neural Information Processing Systems, vol. 30 (2017)
- [5] Staab, R., Vero, M., Balunović, M., Vechev, M.: Beyond memorization: Violating privacy via inference with large language models. In: Proceedings of the Twelfth International Conference on Learning Representations (ICLR) (2024)
- [6] Koto, F., Li, H., Shatnawi, S., Duan, J., Abdelali, A., Mubarak, H., Al-Khalifa, H., Baldwin, T.: ArabicMMLU: Assessing massive multitask language understanding in Arabic. In: Findings of the Association for Computational Linguistics: ACL 2024, pp. 7038–7052 (2024)
- [7] Khondaker, M.T.I., Waheed, A., Nagoudi, E.M.B., Abdul-Mageed, M.: GPTAraEval: A comprehensive evaluation of ChatGPT on Arabic NLP. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Singapore, pp. 220–247 (2023)
- [8] Abdelali, A., Mubarak, H., Chowdhury, S.A., Hasanain, M., Mousi, B., Boughorbel, S., Abdaljalil, S., El Kheir, Y., Izham, D., Dalvi, F., Hawasly, M., Nazar, N., Elshahawy, Y., Ali, A., Durrani, N., Milic-Frayling, N., Alam, F.: LAraBench: Benchmarking Arabic AI with large language models. In: Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers), St. Julian’s, Malta, pp. 487–520 (2024)
- [9] Gallegos, I.O., Rossi, R.A., Barrow, J., Tanjim, M.M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., Ahmed, N.K.: Bias and fairness in large language models: A survey. *Computational Linguistics* **50**(3), 1097–1179 (2024)
- [10] Kotek, H., Dockum, R., Sun, D.: Gender bias and stereotypes in large language models. In: Proceedings of the ACM Collective Intelligence Conference, pp. 12–24 (2023)
- [11] Aly, R., Allam, Y., Gaber, R., Basta, C.: ArGAN: Arabic gender, ability, and nationality dataset for evaluating biases in large language models. In: Proceedings of the 6th Workshop on Gender Bias in Natural Language Processing (GeBNLP), Vienna, Austria (2025)
- [12] Kadaoui, K., Magdy, S.M., Waheed, A., Khondaker, M.T.I., El-Shangiti, A.O.,

- Nagoudi, E.M.B., Abdul-Mageed, M.: TARJAMAT: Evaluation of Bard and ChatGPT on machine translation of ten Arabic varieties. In: Proceedings of the First Arabic Natural Language Processing Conference (ArabicNLP 2023), Singapore, pp. 52–75 (2023)
- [13] Lyu, H., Huang, J., Zhang, D., Yu, Y., Mou, X., Pan, J., Yang, Z., Wei, Z., Luo, J.: GPT-4V(ision) as a social media analysis engine. arXiv preprint arXiv:2311.07547 (2023)
- [14] Rao, D., Yarowsky, D., Shreevats, A., Gupta, M.: Classifying latent user attributes in Twitter. In: Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents, pp. 37–44 (2010)
- [15] Burger, J.D., Henderson, J., Kim, G., Zarrella, G.: Discriminating gender on Twitter. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp. 1301–1309 (2011)
- [16] Nguyen, D., Trieschnigg, D., Dogruöz, A.S., Gravel, R., Theune, M., Meder, T., Jong, F.: Computational sociolinguistics: A survey. *Computational Linguistics* **42**(3), 537–593 (2016)
- [17] Tellez, E.S., Miranda-Jiménez, S., Moctezuma, D., Graff, M., Salgado, V., Ortiz-Bejar, J.: Gender identification through multi-modal tweet analysis using MicroTC and bag of visual words. In: CLEF 2018 Labs and Workshops, Notebook Papers (2018)
- [18] Daneshvar, S., Inkpen, D.: Gender identification in Twitter using N-grams and LSA: Notebook for PAN at CLEF 2018. In: CLEF 2018 Labs and Workshops, Notebook Papers (2018)
- [19] Zhang, C., Abdul-Mageed, M.: BERT-based arabic social media author profiling. arXiv preprint arXiv:1909.04181 (2019)
- [20] Bsir, B., Abraham, A., Zrigui, M.: Transformer models for authorship profiling in Arabic social media texts. *International Journal of Computer Information Systems and Industrial Management Applications* **16**, 623–635 (2024)
- [21] Antoun, W., Baly, F., Hajj, H.: AraBERT: Transformer-based model for Arabic language understanding. In: Proceedings of the LREC 2020 Workshop on Language Resources and Evaluation Conference, pp. 9–15 (2020)
- [22] Abdul-Mageed, M., Elmadany, A., Nagoudi, E.M.B.: ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 7088–7105 (2021)

- [23] Alshaalan, R., Al-Khalifa, H.: Advancing author gender identification in modern standard arabic with innovative deep learning and textual feature techniques. *Information* **15**(12), 779 (2024)
- [24] Alsadhan, N.A.: Is AI catching up to human expression? exploring emotion, personality, authorship, and linguistic style in English and Arabic with six large language models. arXiv preprint arXiv:2603.23251 (2026)
- [25] Lafargue, V., Guerra-Adames, A., Claeys, E., Vuichard, E., Loubes, J.-M.: Probing cultural signals in large language models through author profiling. arXiv preprint arXiv:2603.16749 (2026)
- [26] Al-Sabahi, K., Balhaf, A.F.M., Al Shamy, M.M.Z., Al-Qatf, M., Yang, K., Al-Badwi, A.: Benchmarking large language models on Arabic parsing. *International Journal of Information Management Data Insights* **6**(1), 100404 (2026)
- [27] Al Kautsar, M.D., Almheiri, S., Ahsan, M., Elbouardi, B., Samih, Y., Ahmad, S., Keleg, A., El Herraoui, O., Elzekey, K., Freihat, A.A., Anwar, M., Xie, Z., Liang, J., Al Nasar, M.R., Nakov, P., Koto, F.: Cultural benchmarking of LLMs in standard and dialectal Arabic dialogues. arXiv preprint arXiv:2605.00119 (2026)
- [28] Sengupta, N., Sahu, S.K., Jia, B., Katipomu, S., Li, H., Koto, F., Marshall, W., Gosal, G., Liu, C., Chen, Z., et al.: Jais and Jais-Chat: Arabic-centric foundation and instruction-tuned open generative large language models. arXiv preprint arXiv:2308.16149 (2023)
- [29] Anwar, M., Freihat, A., Ibrahim, G., Awad, M., Sadallah, A., Gosal, G., et al.: Jais 2: A family of Arabic-centric open large language models. Technical report, Institute of Foundation Models, MBZUAI (2025). Model card: <https://huggingface.co/inceptionai/Jais-2-8B-Chat>
- [30] Alzubaidi, A., Alsuwaidi, S., Boussaha, B.E.A., AlQadi, L., Alkaabi, O., Alyafeai, M., Alobeidli, H., Hacid, H.: Evaluating Arabic large language models: A survey of benchmarks, methods, and gaps. arXiv preprint arXiv:2510.13430 (2025)
- [31] Bolukbasi, T., Chang, K.-W., Zou, J.Y., Saligrama, V., Kalai, A.T.: Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In: *Advances in Neural Information Processing Systems*, vol. 29, pp. 4349–4357 (2016)
- [32] Blodgett, S.L., Barocas, S., Daumé III, H., Wallach, H.: Language (technology) is power: A critical survey of “bias” in NLP. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 5454–5476 (2020)
- [33] Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.-W.: Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language*

Processing, pp. 2979–2989 (2017)

- [34] El-Haj, M.: ArabJobs: A multinational corpus of Arabic job ads. In: Proceedings of ArabicNLP 2025 (2025)
- [35] Alhafni, B., Habash, N., Bouamor, H.: User-centric gender rewriting. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 618–631. Association for Computational Linguistics, Seattle, United States (2022)
- [36] Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., et al.: Language models (mostly) know what they know. arXiv preprint arXiv:2207.05221 (2022)
- [37] Gemini Team, Google: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
- [38] OpenAI: GPT-5.4 mini Model Documentation. OpenAI API Documentation. <https://developers.openai.com/api/docs/models/gpt-5.4-mini>. Accessed July 2026 (2026)
- [39] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al.: Transformers: State-of-the-art natural language processing. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pp. 38–45 (2020)
- [40] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al.: Language models are few-shot learners. In: Advances in Neural Information Processing Systems, vol. 33, pp. 1877–1901 (2020)
- [41] Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V., Zhou, D., et al.: Chain-of-thought prompting elicits reasoning in large language models. In: Advances in Neural Information Processing Systems, vol. 35, pp. 24824–24837 (2022)
- [42] McNemar, Q.: Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika* **12**(2), 153–157 (1947)
- [43] AlQuran Cloud: The Holy Quran – Arabic Text. <https://api.alquran.cloud>. Accessed July 2025 (2025)
- [44] gurgutan: Sunnah Arabic-English Dataset: 14 Books of Authenticated Hadiths. HuggingFace Datasets. https://huggingface.co/datasets/gurgutan/sunnah_ar_en_dataset. Accessed July 2025 (2025)

- [45] Habash, N.Y.: Introduction to Arabic Natural Language Processing. Morgan & Claypool Publishers, San Rafael, CA (2010)
- [46] Mashaabi, M., Al-Khalifa, S., Al-Khalifa, H.: A survey of large language models for Arabic language and its dialects. arXiv preprint arXiv:2410.20238 (2024)
- [47] Mousi, B., Durrani, N., Ahmad, F., Hasan, M.A., Hasanain, M., Kabbani, T., Dalvi, F., Chowdhury, S.A., Alam, F.: AraDiCE: Benchmarks for dialectal and cultural capabilities in LLMs. In: Proceedings of the 31st International Conference on Computational Linguistics, Abu Dhabi, UAE, pp. 4186–4218 (2025)