Arabic Annotation Is Not Binary:

LabelBench, a Reproducible Audit of Automated Data Labeling

Bayanat Labs Research

bayanatlabs.com

Technical Report v0.1, revised July 2026

Abstract

Language models are increasingly proposed as substitutes for human data annotators, but aggregate accuracy on simple tasks can hide failure on the taxonomies, dialects, and risk thresholds that make production Arabic labeling difficult. We introduce LabelBench, a checksummed, resumable audit harness for Arabic annotation. The first release covers binary toxicity, seven-way multi-label toxicity, 60-way intent classification, and five-region dialect identification. It reports exact-match accuracy, macro-F1, paired bootstrap intervals, calibration diagnostics, source-leakage audits, and a conservative selective-automation diagnostic: the fraction of items that can be accepted while a Wilson lower bound remains above a target accuracy.

The headline finding is an annotation complexity cliff. On identical held-out samples, Qwen 2.5 3B reaches 78.0% accuracy on binary toxicity but only 25.3% exact match on the full toxicity taxonomy. Arabic-specialized Fanar 1 9B improves those scores to 88.0% and 31.3%, respectively, while a small character n-gram supervised diagnostic reaches 97.0% and 49.0%. On a balanced native dialect set, Qwen 2.5 3B reaches 22.4% accuracy, near the 20% chance floor, and predicts Egyptian for 70 to 80% of every non-Egyptian region. Fanar improves dialect accuracy to 47.8%, but remains 22.0 points behind a 500-example character diagnostic. Because the v0.1 generative runs emit categorical JSON rather than calibrated confidence, their zero selective coverage is a harness limitation, not evidence that selective automation is impossible. The contribution is therefore not a state-of-the-art model ranking. It is a reproducible Arabic audit protocol showing how taxonomy, dialect, confidence, and human review must be measured before automated labels are trusted.

1 Introduction

The commercial promise of model-assisted annotation is straightforward: a model proposes labels, humans review only uncertain cases, and cost falls without sacrificing quality. For Arabic, that promise is tempting and dangerous. Arabic production queues are not one language in one register. They mix dialects, countries, scripts, code-switching, profanity conventions, domain-specific policies, and label taxonomies whose operational meaning depends on fine distinctions. A model that looks competent on a binary proxy task can still fail at the decision actually needed by a moderation queue, a retrieval evaluator, a speech system, or a post-training data engine.

This paper asks a narrower and more actionable question than "Can models label Arabic?" We ask:

1. When the label space becomes more realistic, how quickly does zero-shot automated labeling degrade?
2. Does an Arabic-specialized local model reduce the failure modes of a generic local instruction model?
3. Do the evaluated outputs include enough calibrated confidence signal to justify auto-accepting any portion of the queue?
4. Can a laptop-scale benchmark expose these issues reproducibly without redistributing restricted text?

The answer is sobering but useful. Specialization helps: Fanar 1 9B beats Qwen 2.5 3B by 10.0 points on binary toxicity, 6.0 points on multi-label exact match, and 25.4 points on dialect identification. However, specialization does not, under this zero-shot JSON-only protocol, provide a calibrated auto-accept policy at 95% accuracy. A small supervised diagnostic still outperforms the evaluated zero-shot systems on all paired tasks. The narrow lesson is that Arabic labeling automation requires task-specific measurement, calibrated confidence, and native review rather than generic "AI annotator" claims.

2 Contributions

We make five contributions.

1. **A reproducible Arabic annotation audit.** LabelBench pins source revisions, raw-file checksums, normalized manifest hashes, model digests, prompts, runtime metadata, and append-only prediction logs.
2. **A production-oriented selective-automation diagnostic.** SAC is a simple Wilson-bound coverage calculation connected to selective classification, calibration, and conformal risk-control work [4, 9, 11].
3. **A quantified source-confound audit.** We show that combined Arabic dialect corpora can let models learn corpus provenance rather than dialect; a trivial source-majority shortcut would reach 80.1% on combined IADD.
4. **Paired evidence for an annotation complexity cliff.** The same held-out items are used for Qwen, Fa-

nar, and supervised diagnostics, enabling paired bootstrap intervals and exact McNemar tests.

5. **A human adjudication path.** The release includes a blinded, right-to-left, offline review bundle so future versions can separate model errors from gold ambiguity and taxonomy defects.

3 Scope of the claim

The paper is intentionally framed as an audit report rather than a leaderboard. This distinction matters. A leaderboard asks which system is best under a fixed benchmark. An annotation audit asks whether a system is safe to insert into a labeling workflow with a stated quality target. The latter requires item-level reproducibility, confidence calibration, gold-label scrutiny, and an escalation rule for uncertain examples. A model can improve average accuracy and still be unusable for automatic acceptance if its confidence ranking does not separate correct from incorrect labels. Conversely, a model with imperfect average accuracy can be operationally useful if a small, well-calibrated slice can be accepted and the rest escalated.

LabelBench v0.1 therefore supports three limited claims. First, the selected Arabic manifests expose a large gap between binary proxies and production-like taxonomies. Second, the evaluated laptop-scale zero-shot systems are not validated automation policies under the current prompting and confidence protocol. Third, the harness is sufficient to run stronger tests: frontier models, few-shot/native prompts, human adjudication, and risk-controlled selection can be added without changing the provenance or logging contract. It does not support a general claim that all generative models fail at Arabic annotation, nor a claim that character n-gram diagnostics are state of the art.

4 Related work

Arabic resources and shared tasks. AraTox provides a multi-dialect, multi-label Arabic toxicity benchmark with expert labels across six Arabic varieties [3]. MASSIVE provides localized virtual-assistant utterances across many languages, including ar-SA, and is useful as a high-structure intent classification contrast [8]. IADD integrates multiple dialect identification datasets [17]; DART is a crowd-labeled Arabic dialect tweet corpus covering the five broad regions used here [2]. LabelBench is not a replacement for community benchmarks: NADI is the canonical Arabic dialect-ID shared-task series and now includes multi-label country-level dialect identification and dialectness estimation [1]. OSACT5 is the relevant shared-task lineage for Arabic offensive language and fine-grained hate-speech detection [13]. Future LabelBench releases should add NADI/OSACT-compatible tasks so results are directly comparable to the field.

Dialect ambiguity and dataset shortcuts. Arabic dialect identification is not a clean single-label problem. Keleg and Magdy show that many apparent system errors reflect incomplete dialect labels and argue for multi-label framing [12]. This motivates our human adjudication protocol and limits how strongly any DART-only five-region result should be interpreted.

Annotation quality. The data-labeling literature has long shown that cheap annotation can be useful when disagreement, worker quality, and aggregation are measured rather than assumed [7, 14]. LabelBench adopts the same posture toward automated annotators: a model prediction is not a label until its quality is measured under the relevant taxonomy and review policy. Recent LLM-annotation work shows that frontier models can sometimes match or exceed crowd annotators on selected text tasks [10], that zero-shot LLMs can assist computational social-science annotation while still lagging fine-tuned classifiers on taxonomic labeling [18], and that off-the-shelf multilingual LLM annotation varies substantially across tasks and low-resource languages [5]. The Alternative Annotator Test provides a statistical procedure for deciding when an LLM can replace human annotations on a task-specific sample [6]. LabelBench should be read as an audit harness that can feed such procedures, not as a substitute for them.

Selective prediction and calibration. Selective classification evaluates the tradeoff between coverage and risk: a system may abstain when it is likely to be wrong [9]. Calibration asks whether stated confidence reflects empirical correctness [11]. In annotation operations, these are not academic extras. They determine whether a queue can be partially automated without quietly corrupting training data. SAC is a deliberately simple launch diagnostic; conformal risk control and selective conformal risk control provide stronger finite-sample frameworks for future releases [4, 16].

5 Benchmark design

5.1 Tasks and data

Table 1 summarizes LabelBench v0.1. AraTox contains approximately 36,700 expert-annotated comments spanning six Arabic varieties and seven toxicity categories [3]. We derive a binary toxicity view and retain the original multi-label view. MASSIVE ar-SA is included as a translated, high-structure intent benchmark rather than treated as native dialect data. For dialect identification we isolate only the DART source inside IADD, for the reasons in Section 5.2.

Table 1: LabelBench v0.1 tasks. Train counts apply only to supervised diagnostics; zero-shot models receive no demonstrations.

Task	Source	Native / localized	Labels	Train	Test
Binary toxicity	AraTox	native, multi-dialect	2	33,038	3,670
Multi-label toxicity	AraTox	native, multi-dialect	7	33,038	3,670
Intent classification	MASSIVE ar-SA	localized	60	11,514	2,974
Dialect region	IADD/DART only	native, crowd-labeled	5	500	500

Table 2: IADD source-region confounding. Counts are examples in the five regions evaluated here. The source-majority shortcut would score 80.1%.

Source	EGY	GLF	IRQ	LEV	MGH	Total
AOC	4414	6259	9	6503	28	17213
DART	423	423	207	397	331	1781
PADIC	0	0	0	14426	21639	36065
SHAMI	0	0	0	66247	0	66247
TSAC	0	0	0	0	11998	11998

5.2 Why source confounding matters

Dialect identification datasets often combine corpora collected for different projects. If one corpus contributes mostly Levantine text and another mostly Maghrebi text, a classifier can learn collection artifacts, topic patterns, or annotation provenance instead of dialect. IADD is valuable because it integrates several resources, but that integration also creates a serious evaluation trap.

Table 2 quantifies the issue. A trivial rule that predicts each IADD source's majority region is 80.1% accurate over the five target regions. This number is higher than every dialect result in our DART-only evaluation, including the supervised diagnostic. Reporting such a score as dialect competence would be misleading. We therefore evaluate only DART, which contains all five target regions under one source. After normalization, DART yields 1,761 clean unique rows: 420 Gulf, 419 Egyptian, 391 Levantine, 328 Maghrebi, and 203 Iraqi. We draw a balanced 100-per-region evaluation set and a disjoint 100-per-region diagnostic training set.

5.3 Integrity controls

Every network source is pinned to a content revision and SHA-256 digest. Every normalized train, validation, test, and sampled manifest has a second digest. Prediction rows are appended atomically and keyed by stable item ID; resumed runs validate cached gold labels and reject unknown or duplicate IDs. Reports record exact local model digests, prompt and completion counts where available, runtime, Python version, code revision, and prediction-file digest.

We audit normalized train/test overlap before scoring. MASSIVE ar-SA contains 8.41% normalized overlap and 63 conflicting overlaps; the character baseline falls from 74.24% to 73.27% after overlapped test items are removed. AraTox binary overlap is 0.49% with no binary-label conflicts and no material score change. We report these findings rather than silently deduplicating a canonical test set.

5.4 Models

The generative baselines use locally quantized instruction models through Ollama with temperature zero, a fixed seed, and strict JSON outputs. Qwen 2.5 3B is the generic laptop-scale principal baseline. Fanar 1 9B is the Arabic-specialized local model. Qwen 2 7B and Llama 3 8B are included on the dialect task to check whether the Qwen 2.5 dialect collapse is merely a small model artifact. Structured-output validity is scored rather than repaired. These are laptop-scale baselines, not the systems a well-funded production team would necessarily deploy. Frontier and larger Arabic-specialist models remain necessary before making broad model-capability claims.

Supervised diagnostics are intentionally simple: a character 2 to 5-gram multinomial naive Bayes classifier for single-label tasks and a one-vs-rest variant for multi-label toxicity. Multi-label thresholds are selected on a 15% partition of the training split, then the models are refit on all training data. These diagnostics are not state-of-the-art claims. They answer a narrower question: does the held-out manifest contain learnable signal under a small-compute regime?

5.5 Prompting and confidence protocol

The local generative runs use a single zero-shot instruction template per task. The prompt names the closed label set, asks for strict JSON, and does not include demonstrations, rationales, chain-of-thought, or Arabic-native prompt variants. Decoding is deterministic. This design makes the first release cheap, reproducible, and easy to inspect, but it also deliberately under-tests what a production annotation team would try. Few-shot examples, Arabic prompts, self-consistency, and model-specific output calibration are part of the next experiment matrix.

The original v0.1 runs store categorical labels with schema-level confidence set to 1.0. That value records parser validity, not model belief. The current code adds a verbalized-confidence mode for future local runs and an OpenAI Responses adapter that can require a confidence field in the JSON schema. Those extensions are

not retroactively applied to the v0.1 numbers. They are included so that the next release can separate model accuracy from the routing question: which items, if any, can be accepted automatically?

6 Metrics

For single-label tasks we report accuracy, macro-F1, a 1,000-replicate bootstrap 95% interval for accuracy, Brier score, expected calibration error, and a confusion matrix. For multi-label toxicity the primary metric is subset accuracy: a prediction is correct only when its entire label set matches. Micro-F1, macro-F1, Hamming loss, label cardinality, and per-label F1 provide secondary views.

Safe Automation Coverage. For predictions sorted by confidence, consider every threshold without splitting ties. At each threshold, let k be correct among n accepted predictions and let $L(k, n)$ be the 95% Wilson lower confidence bound [15]. For quality target τ,

$$\mathrm{SAC}_\tau = \max_t \frac{n_t}{N} \quad \text{s.t.} \quad L(k_t, n_t) \geq \tau.$$

We report SAC95 and SAC98 as an operational diagnostic, not as a new statistical guarantee. Empirical accuracy alone is insufficient when the accepted set is small, but selecting a threshold on the same evaluation set is optimistic. The code now supports calibration-split SAC, where the threshold is chosen on a calibration partition and evaluated separately. A future benchmark release should replace or accompany SAC with conformal risk-control methods. Models that emit only schema-level confidence of 1.0 cannot manufacture a useful ranking and therefore must satisfy the bound on the whole tied set.

Paired tests. When systems share the same evaluation items, we report paired bootstrap intervals for accuracy or exact-match differences and exact two-sided McNemar tests on the correctness contingency table. The paired design is important: it distinguishes a real model difference from an easier or harder sample. For the eight central v0.1 paired comparisons, the release JSON also reports Holm-adjusted McNemar p-values. Effects that do not survive this correction are treated as exploratory, even when their raw paired tests are below 0.05.

7 Results

Figure 1 gives the central result. Qwen's 78.0% binary toxicity accuracy drops to 25.3% when exact agreement with the seven-label taxonomy is required. Fanar improves the corresponding scores to 88.0% and 31.3%, but does not close the gap to small supervised diagnostics at 97.0% and 49.0%. For five-region dialect identification, Qwen scores 22.4% accuracy and 14.9% macro-F1 against a 20% chance floor. Fanar reaches 47.8%, while a 500-example character diagnostic reaches 69.8% accuracy and 68.8% macro-F1.

7.1 Binary toxicity is not enough

Binary toxicity is the easiest and most forgiving task in the release. On the 100-item paired slice, Qwen reaches 78.0% accuracy. Fanar reaches 88.0%, with only one false positive but eleven toxic items missed as non-toxic. The character diagnostic reaches 97.0%. On the full AraTox binary test set, the character diagnostic reaches 96.27% accuracy and 95.47% macro-F1, while a majority baseline is 71.88% accurate but only 41.82% macro-F1. This confirms that the binary label is learnable and that raw accuracy can be inflated by class imbalance.

7.2 Multi-label toxicity exposes category errors

The multi-label task is operationally closer to a real policy queue. A system must choose among appearance attacks, cussing, hatred, racial attacks, sexual content, violence, and the mutually exclusive "not" label. On the 300-item paired sample, Qwen's exact match is 25.3% and macro-F1 is 39.3%. Fanar improves to 31.3% exact match and 50.5% macro-F1. The tuned character OVR diagnostic reaches 49.0% exact match and 73.2% macro-F1 on the same sample. On the full AraTox test set, the tuned diagnostic reaches 53.73% exact match, 73.35% macro-F1, 74.70% micro-F1, and 11.64% Hamming loss.

Table 4 shows that the failure is not merely low exact match. Qwen is particularly weak on violence, with 10.3% F1. Fanar raises violence F1 to 42.3% and sexual-content F1 to 71.8%, but drops appearance F1 to 20.0%. The average number of predicted labels is close to the gold average for Qwen (1.40 predicted versus 1.38 gold), so the problem is not only under-prediction. The model often chooses the wrong categories.

7.3 Dialect collapse is systematic

Qwen 2.5 predicts Egyptian for 85% of Egyptian items, but also for 75% of Gulf, 70% of Iraqi, 80% of Levantine, and 70% of Maghrebi items. Recall is 4% for Gulf, 6% for Iraqi, 13% for Levantine, and 4% for Maghrebi. Qwen 2 7B changes the dominant prediction from Egyptian toward Levantine but reaches nearly identical accuracy. Llama 3 reaches 24.6%. This indicates that the failure is not just one small model being noisy.

Fanar 1 9B reaches 47.8%, with 60% Egyptian, 75% Gulf, 13% Iraqi, 50% Levantine, and 41% Maghrebi recall. Regional specialization substantially reduces collapse, although Iraqi remains a severe weakness and the supervised diagnostic remains 22 points higher. The character diagnostic's recalls are 81% Egyptian, 70% Gulf, 88% Iraqi, 70% Levantine, and 40% Maghrebi. Maghrebi remains difficult even for the diagnostic, consistent with

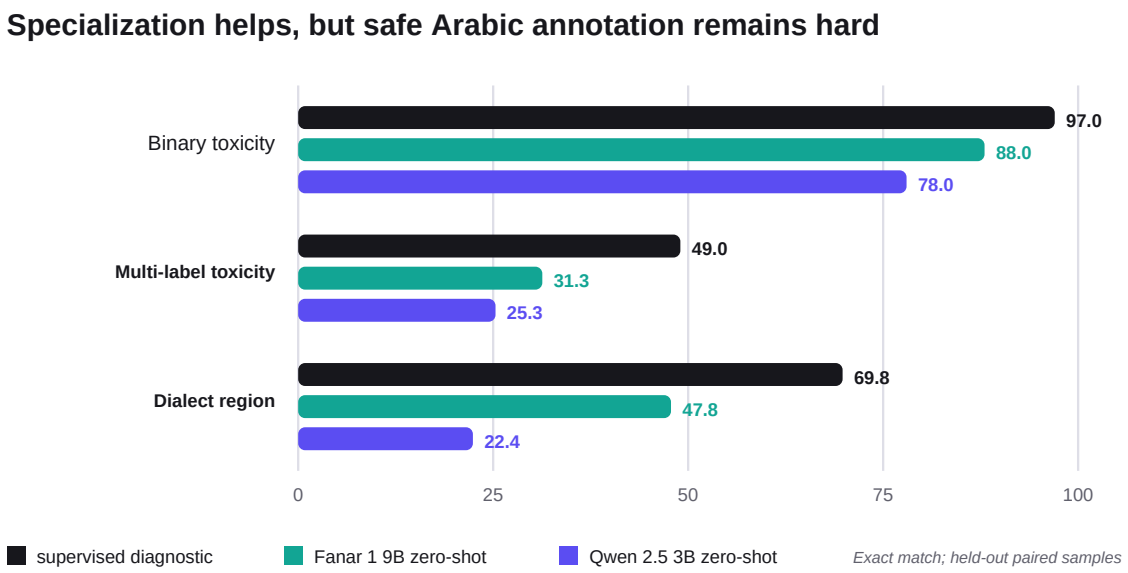

Figure 1: The annotation complexity cliff on identical held-out manifests. Binary and multi-label toxicity use the same seeded AraTox samples for all three systems. The dialect diagnostic is trained on 500 DART items disjoint from the balanced 500-item evaluation manifest.

Table 3: Held-out results. Exact is accuracy for single-label tasks and subset accuracy for multi-label toxicity. SAC95 is the maximum coverage whose Wilson lower bound remains at least 95%.

Task	Model	Items	Exact	Macro-F1	Micro-F1	SAC95
Binary toxicity	Qwen 2.5 3B	100	78.0	69.8	–	0.0
Binary toxicity	Fanar 1 9B	100	88.0	83.1	–	0.0
Binary toxicity	Char NB	100	**97.0**	**95.0**	–	89.0
Multi-label toxicity	Qwen 2.5 3B	300	25.3	39.3	43.0	0.0
Multi-label toxicity	Fanar 1 9B	300	31.3	50.5	55.0	0.0
Multi-label toxicity	Char OVR	300	**49.0**	**73.2**	**73.9**	0.0
Dialect region	Qwen 2.5 3B	500	22.4	14.9	–	0.0
Dialect region	Qwen 2 7B	500	22.6	16.2	–	0.0
Dialect region	Llama 3 8B	500	24.6	18.7	–	0.0
Dialect region	Fanar 1 9B	500	47.8	47.3	–	0.0
Dialect region	Char NB	500	**69.8**	**68.8**	–	0.0
Intent classification	Char NB	2974	74.2	65.1	–	39.5

DART's reported lower agreement for some dialect labels [2].

7.4 Paired significance tests

All central comparisons use identical items. Exact McNemar p-values below are raw unless otherwise stated; machine-readable results include Holm family-wise corrections across the eight paired comparisons. Fanar exceeds Qwen by 10.0 points on binary toxicity (paired-bootstrap 95% CI: 2.0 to 18.0; raw $p = 0.031$, Holm $p = 0.090$), while the character diagnostic exceeds Fanar by 9.0 points (2.0 to 17.0; raw $p = 0.035$, Holm $p = 0.090$) and Qwen by 19.0 points (11.0 to 28.0; Holm $p = 8.4 \times 10^{-5}$).

On multi-label toxicity, the character diagnostic exceeds Qwen by 23.7 points (17.7 to 29.7; Holm $p < 10^{-12}$). Fanar exceeds Qwen by 6.0 points (1.0 to 11.3; raw $p = 0.030$, Holm $p = 0.090$), while the character diagnostic remains 17.7 points higher than Fanar (11.7 to 23.7; Holm $p < 5 \times 10^{-7}$). On dialect identification, Fanar exceeds Qwen by 25.4 points (20.2 to 30.6; Holm $p < 2 \times 10^{-18}$), while the character diagnostic remains 22.0 points above Fanar (16.2 to 27.6; Holm $p < 2 \times 10^{-12}$).

7.5 Confidence was not elicited, so SAC is diagnostic only

The zero-shot structured classifiers return categorical schema outputs rather than calibrated class probabilities, creating a single confidence tie. None reaches non-zero

Table 4: Per-label F1 on the 300-item multi-label toxicity sample. The weak categories differ by model, which matters for policy routing.

Label	Qwen	Fanar	Char OVR
Appearance	34.7	20.0	72.3
Cussing	26.2	62.9	73.6
Hatred	50.0	47.3	70.5
Racial	36.4	39.0	46.7
Sexual	49.5	71.8	86.9
Violence	10.3	42.3	70.6
Not	68.1	69.9	92.1

SAC95 under this tied-confidence protocol. This should not be read as evidence that generative models cannot support selective automation. It shows that a JSON label alone is insufficient for automation routing. The harness now supports a verbalized-confidence mode for future local-model runs; frontier models should additionally be tested with token log probabilities, self-consistency agreement, or conformal risk-control calibration where available. The tuned multi-label character system also reaches zero SAC95 despite stronger F1, showing that label quality and safe triage are separate problems. The only substantial SAC95 in v0.1 is the binary character diagnostic: 89.0% on the seeded 100-item sample and 100% on the full AraTox binary test. Intent classification reaches 39.5% SAC95, a reminder that selective automation is possible when confidence ranking is useful.

8 Robustness checks

The release is designed to reduce three common benchmark failure modes. First, the DART-only dialect protocol avoids the 80.1% source-majority shortcut in combined IADD. Second, all central model comparisons use paired items, so differences are tested on the same examples rather than across different samples. Third, the smaller paired LLM slices are interpreted alongside full-test supervised diagnostics, which confirm that the binary and multi-label AraTox tasks are learnable under a low-compute text baseline. These checks support a narrow claim: under one zero-shot local-model protocol, including an Arabic-specialized model, these systems do not yet provide a validated Arabic label-automation policy on these taxonomies without measured human review.

9 Human adjudication protocol

Automated scores cannot detect every bad gold label, ambiguous dialectal form, or taxonomy defect. LabelBench therefore includes a blinded, offline, right-to-left review application. A release bundle contains no model prediction. Three qualified native reviewers independently assign the full label set, a three-level confidence judgment, and optional flags for ambiguity, possible bad gold, personally identifying information, and distressing content. The "not" label is enforced as mutually exclusive.

Aggregation validates task IDs and label sets, requires complete submissions, computes unanimous and majority consensus, and surfaces disagreement for adjudication. Reviewer IDs are assigned pseudonyms; progress remains in local browser storage until an explicit JSON export. This protocol is prepared for a 300-item AraTox audit. Human results are intentionally not fabricated in this technical report.

10 What would justify annotator replacement?

The v0.1 evidence is deliberately below the bar for replacing human annotators. A credible replacement or auto-acceptance claim would require at least four additions.

1. **Human reference study.** Run three independent qualified native reviewers on the same blind bundle, report Krippendorff's α or an appropriate multi-label agreement statistic, estimate gold-label error, and adjudicate persistent disagreements before treating the dataset as a replacement target.
2. **Frontier and specialist snapshots.** Evaluate dated snapshots of the systems likely to be deployed: frontier APIs, larger Arabic-specialist models, and a fine-tuned Arabic encoder such as MARBERT-style baselines where licensing permits. Local sub-10B quantized runs are useful smoke tests, not production evidence.
3. **Prompt and calibration matrix.** Cross zero-shot and few-shot prompts, English and Arabic instructions, self-consistency, and confidence elicitation. Selective automation should then be evaluated on a held-out calibration/evaluation split and compared with alt-test or conformal risk-control procedures.
4. **Power and economics.** Increase paired sample sizes before turning small differences into claims; report latency, cost per 10,000 attempted labels, cost per correctly accepted label, and the share escalated to humans.

Under that protocol, the central output would not be a single accuracy table. It would be a decision curve: expected error rate and cost as automation coverage increases, with human review used wherever the quality floor is not statistically supported.

11 Limitations and ethics

This is a launch audit, not a top-tier benchmark paper or ranking of all frontier systems. The principal generative results are local quantized models; closed frontier systems and larger Arabic-specialist models remain untested. The prompts are zero-shot and English-authored; no few-shot, Arabic-prompt, chain-of-thought, self-consistency,

or prompt-sensitivity axis is reported. The headline LLM samples are modest (100, 300, and 500 items), while the supervised diagnostics train in-domain and therefore should be interpreted as learnability controls rather than competitive model baselines.

The human adjudication protocol is implemented but not yet run, so v0.1 cannot estimate gold-label ambiguity, inter-annotator agreement, or human replacement validity. Future releases should add three-rater adjudication with Krippendorff's α, dated frontier snapshots, native Arabic and few-shot prompts, NADI/OSACT-compatible tasks, held-out calibration splits, and an alt-test or conformal risk-control analysis before making annotator-replacement claims.

The dialect evaluation covers broad regions, not countries, cities, register, code-switching, Arabizi, or speech. Dialect is socially variable and cannot be reduced to geography. DART itself reported lower human agreement for Iraqi and Maghrebi labels [2]; model errors in those groups must not be interpreted as deficiencies in the varieties or their speakers.

Toxicity data can harm reviewers. The adjudication interface includes a distressing-content flag, supports offline work, and should be deployed with informed consent, opt-out, breaks, and access to support. Public release must follow each source's terms. LabelBench redistributes no IADD/DART text and publishes only aggregate metrics and cryptographic manifest identifiers.

12 Conclusion

Arabic labeling automation cannot be certified by a binary accuracy number. Across toxicity and dialect tasks, the work becomes sharply harder when the label space reflects production reality. The correct question is not "Can a model label Arabic?" but "Which taxonomy, which variety, at what proven quality floor, and with how much human review?" LabelBench makes that question measurable.

References

[1] Muhammad Abdul-Mageed, Amr Keleg, Abdel-Rahim Elmadany, Chiyu Zhang, Injy Hamed, Walid Magdy, Houda Bouamor, and Nizar Habash. NADI 2024: The fifth nuanced Arabic dialect identification shared task. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages 709–728, 2024. doi: 10.18653/v1/2024.arabicnlp-1.79.

[2] Israa Alsarsour, Esraa Mohamed, Reem Suwaileh, and Tamer Elsayed. Dart: A large dataset of dialectal arabic tweets. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pages 3666–3674, 2018.

[3] Faisal Alshargi, Ahmed Abulohoom, Sane Yagi, Furat Jabr, Leena Lulu, and Ashraf Elnagar. Aratox: a multi-dialect, multi-label arabic dataset and model benchmark for toxicity detection. *Language Resources and Evaluation*, 60(39), 2026. doi: 10.1007/s10579-026-09917-9.

[4] Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=33XGfHLtZg.

[5] Savita Bhat and Vasudeva Varma. Large language models as annotators: A preliminary evaluation for annotating low-resource language content. In *Proceedings of the 4th Workshop on Evaluation and Comparison of NLP Systems*, pages 100–107, 2023. doi: 10.18653/v1/2023.eval4nlp-1.8.

[6] Nitay Calderon, Roi Reichart, and Rotem Dror. The alternative annotator test for LLM-as-a-judge: How to statistically justify replacing human annotators with LLMs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 16051–16081, 2025. doi: 10.18653/v1/2025.acl-long.782.

[7] A. Philip Dawid and Allan M. Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied Statistics*, 28(1):20–28, 1979.

[8] Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. Massive: A 1m-example multilingual natural language understanding dataset with 51 typologically-diverse languages. *arXiv preprint arXiv:2204.08582*, 2022. doi: 10.48550/arXiv.2204.08582.

[9] Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[10] Fabrizio Gilardi, Meysam Alizadeh, and Mael Kubli. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120, 2023. doi: 10.1073/pnas.2305016120.

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

[12] Amr Keleg and Walid Magdy. Arabic dialect identification under scrutiny: Limitations of single-label classification. In *Proceedings of ArabicNLP 2023*, pages 385–398, 2023. doi: 10.18653/v1/2023.arabicnlp-1.31.

[13] Hamdy Mubarak, Hend Al-Khalifa, and Abdulmohsen Al-Thubaity. Overview of OSACT5 shared task on Arabic offensive language and hate speech detection. In *Proceedings of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pages 162–166, 2022.

[14] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, 2008.

[15] Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

[16] Yunpeng Xu, Wenge Guo, and Zhi Wei. Selective conformal risk control. *arXiv preprint arXiv:2512.12844*, 2025.

[17] Jihad Zahir. Iadd: An integrated arabic dialect identification dataset. *Data in Brief*, 40:107777, 2022. doi: 10.1016/j.dib.2021.107777.

[18] Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291, 2024. doi: 10.1162/coli_a_00502.