Arabic Safety Alignment as Selective Refusal: An Empirical Study of SFT, DPO, and Guard Calibration

Mohamad Zbib
American University of Beirut
mbz02@mail.aub.edu

Ammar Mohanna
American University of Beirut
am288@aub.edu.lb

Abstract

Arabic large language models must refuse harmful prompts without over-refusing benign or sensitive prompts, yet a single refusal rate hides this trade-off. We evaluate it using benign refusal B and harmful-prompt refusal H, where H measures refusal rather than harmful compliance. Across five Arabic-capable models and 130 runs on the full human-written AraSafe set, refusal-only supervised fine-tuning (SFT) collapses toward blanket refusal, whereas selected mixed-SFT configurations reach $H \approx 90$–93% at $B = 14$–23%; four selected configurations exceed $H = 90\%$ in all three runs, while Fanar does so in two of three. Direct Preference Optimization (DPO) and inference guards change B and H differently across models rather than acting as uniform upgrades. In a blinded 300-response audit, annotator binary-refusal agreement is 89.0% ($\kappa = 0.78$); Qwen3Guard and Aya Expanse 32B reach 88.7% and 91.0% accuracy, respectively, with no conclusive paired difference. Selected SFT raises H on Arabizi for all five models, but none reaches 90%, showing only partial transfer from Modern Standard Arabic; overall, the results support model-specific operating-point selection: set a deployment target and retain only interventions that improve it.

1 Introduction

Arabic safety is not a single input space. Users write in Modern Standard Arabic (MSA), regional dialects such as Egyptian and Levantine, romanized Arabizi, and noisy text, and many benign prompts are sensitive: news, medical, legal, and policy questions that share vocabulary with genuinely harmful requests (Ashraf et al., 2025; Mousi et al., 2025). A safety method for Arabic must therefore refuse harmful prompts while staying usable across these forms. A single refusal rate cannot capture this requirement because a model can appear protective simply by refusing everything. The practical question is whether standard alignment interventions improve this trade-off consistently across Arabic-capable models and writing forms.

We evaluate selective refusal with two rates: benign refusal $B = P(\text{refusal} \mid \text{benign prompt})$ and harmful-prompt refusal $H = P(\text{refusal} \mid \text{harmful prompt})$, measured on the human-written AraSafe benchmark (Mubarak et al., 2025). Crucially, H records whether a response refuses; its complement is not automatically harmful compliance. A useful intervention must lower B or raise H without silently damaging the other axis, so every method is judged by movement on the (B, H) plane rather than either rate alone (Röttger et al., 2024; Cui et al., 2024).

Figure 1 motivates the study: five Arabic-capable models begin at different measured locations in the (B, H) plane. This variation motivates model-by-model testing; it does not establish universal model categories or transferable intervention rules. Our central claim is narrower: *Arabic selective refusal is a model-specific operating-point selection problem: measure benign refusal B and harmful-prompt refusal H, specify the deployment target, and retain an intervention only when it improves that target.*

Contributions.

- **Audited evaluation.** We audit B/H labels with 300 blinded responses and a cross-family judge.
- **Model-specific SFT trade-offs.** Refusal-only SFT collapses, while selected mixed-SFT configurations reach $H \approx 90$–93% at $B = 14$–23%.
- **No uniform post-training upgrade.** DPO and guards move models differently; ordering is a null result.
- **Partial cross-script transfer.** Selected SFT raises Arabizi H for every model, but none reaches 90%.

Proceedings of The Fourth Arabic Natural Language Processing Conference (ArabicNLP 2026), pages 228–246
October 28-29, 2026 ©2026 Association for Computational Linguistics

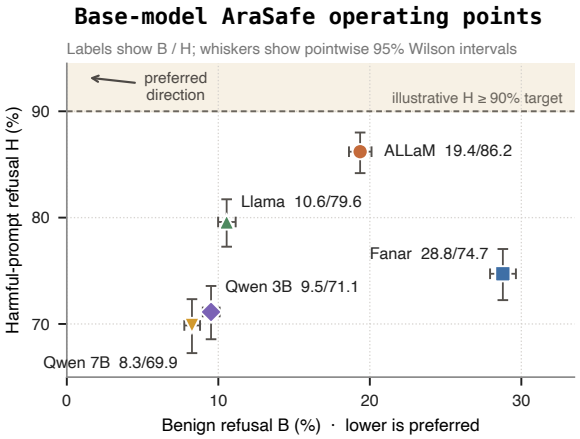

Figure 1: Base-model operating points on AraSafe. Lower benign refusal B and higher harmful-prompt refusal H are preferred; labels report B/H. Whiskers show marginal 95% Wilson intervals for prompt sampling conditional on the automatic judge labels ($n_B = 10{,}823$, $n_H = 1{,}254$), reconstructed from retained rates. The dashed line marks the illustrative $H \geq 90\%$ target, not a universal safety standard.

2 Measuring and Validating Selective Refusal

Metrics and operating points. An operating point is one measured pair (B, H). One tested point dominates another when it has lower or equal B and higher or equal H, with at least one strict inequality; the tested frontier is the non-dominated subset of observed points. We use $H \geq 90\%$ as an illustrative selection constraint and choose the lowest-B feasible candidate. This is not a universal safety standard: a deployment owner must set the target from its own costs, and we report sensitivity to 85%, 90%, and 95% in Table 10.

Evaluation set and judge. We evaluate on the full human-written portion of AraSafe (Mubarak et al., 2025): 12,077 prompts, of which 10,823 are benign and 1,254 harmful. We reserve these prompts for evaluation; none appears in training, direction extraction, or calibration. Because AraSafe is MSA-heavy, we treat the operating points in Sections 3 and 4 as MSA measurements and stress-test them across forms in Section 5. Qwen3Guard-Gen-4B (Zhao et al., 2025; Souly et al., 2024) supplies the binary refusal label.

Human and cross-judge validation. We audited a blinded, stratified sample of 300 responses with two independent annotators and adjudication. Binary-refusal agreement was 89.0% ($\kappa = 0.78$). Against adjudicated labels, Qwen3Guard achieved 88.7% accuracy and 0.887 macro-F1, with a clustered-bootstrap 95% accuracy interval of [85.3, 91.7]; Aya Expanse 32B (Dang et al., 2024) achieved 91.0% accuracy and 0.910 macro-F1. Their paired difference was inconclusive (McNemar $p = 0.382$; Qwen-minus-Aya macro-F1 difference -2.3 points, 95% interval $[-6.4, 2.0]$). Qwen3Guard accuracy was 86.7% on Qwen-family outputs and 90.0% otherwise; the gap interval $[-3.3, 10.3]$ neither establishes nor excludes family bias. Reported form-specific accuracies range from 87.5% to 95.0% on MSA, Egyptian, Levantine, and noisy Arabic subsets, but fall to 77.5% on Arabizi, where macro-F1 is 0.763. Full audit statistics appear in Table 7.

Models. The five base models span sizes and providers: Qwen2.5-3B and Qwen2.5-7B Instruct, Meta-Llama-3-8B Instruct, Fanar-1-9B, and ALLaM-7B Instruct. Two are Arabic-centric and three are multilingual; neither provenance group occupies one common starting location in Figure 1. Full identifiers are in Appendix F.

Training and preference data. Harmful SFT and preference prompts come from an MSA translation of BeaverTails (Ji et al., 2023), translated with the Hala English–Arabic models (Hammoud et al., 2025); benign SFT examples are separately drawn from Hala-4.6M-SFT. DPO uses 10K benign and 10K harmful preference pairs: harmful pairs rank a refusal over a harmful completion, while benign pairs rank either two helpful answers (V1) or a helpful answer over a refusal (V2) from the Arabic Data Is Better Together collection (Data Is Better Together, 2024). Because the harmful supervision is translated MSA, cross-script transfer is an empirical question rather than an assumption.

Uncertainty and seeds. Wilson intervals quantify prompt-sampling uncertainty conditional on the automatic labels; they do not include judge measurement error, which the audit exposes separately. With 10,823 benign and 1,254 harmful prompts, B half-widths are below one point at the reported rates; H half-widths are about 1.5 points near the selected $H \approx 90\%$ points and 1.8–2.6 points at the base points in Figure 1. Main-text rates are rounded to one decimal place, while appendix tables retain the available source precision; a rounded boundary value is not treated as stable evidence of threshold satisfaction. Primary sweeps use seed 42. Selected mixed-SFT configurations were repeated

Table 1: Notation and scope. Refusal is measured independently on benign and harmful prompts; H is not a harmful-compliance rate.

Term	Meaning
B	$P(\text{refusal} \mid \text{benign prompt})$; false-positive refusal rate
H	$P(\text{refusal} \mid \text{harmful prompt})$; refusal, not harmful compliance
Operating point	One measured pair (B, H) for a model and intervention
Tested frontier	Non-dominated subset of the operating points actually evaluated

Table 2: Data roles in the operating-point search. Evaluation prompts are held out from training and calibration; the remaining stages use the sources and splits listed below.

Split	Used for	Why it matters
AraSafe	Main B/H evaluation	Measures Arabic selective refusal
SFT and DPO data	Training	Supplies helpful and refusal pressure
Boundary set	Stress test only	Checks dialect, Arabizi, noise, and sensitive benign generalization
Activation split	Direction extraction	Learns candidate linear refusal directions
Calibration split	Threshold tuning	Selects per-model thresholds separately from extraction

with seeds 43 and 44: across models, standard deviations span 0.13–0.57 points for B and 0.41–1.10 for H. ALLaM, Llama, Qwen 3B, and Qwen 7B exceed $H = 90\%$ in all three runs; Fanar does so in two of three, with mean $H = 89.95 \pm 0.41$, and is therefore threshold-sensitive. Direction AUC intervals use the Hanley–McNeil estimator.

Scope of the starting points. The five observed base points differ substantially: ALLaM starts with comparatively high H, Fanar with the highest B, and the Qwen and Llama models with lower B but also lower H. These are measured operating points, not discovered clusters, causal diagnoses, or intervention prescriptions. We therefore search candidate operating points separately for each model. Table 1 makes the paper's notation and inferential scope explicit, while Table 2 records which data touch each stage.

3 SFT Candidate Search

Refusal-only SFT collapses toward blanket refusal. We use refusal-only training (Qi et al., 2024) as a diagnostic. As Figure 2 shows, harmful-prompt refusal H saturates within a few steps, while benign refusal B rises toward total refusal: ALLaM, Fanar, and Llama cross $B = 80\%$ by step 10, and both Qwen models by step 30. Thus, refusal pressure must be bounded by benign behavior.

Mixed SFT exposes local candidate points. Adding helpful examples recovers selectivity, but the ratio is a model-specific search (Figure 3). For example, ALLaM moves from $28.1/94.7$ at 70/30 benign/refusal to $17.7/92.4$ at 95/5, whereas Fanar falls to $8.6/79.0$ at 95/5 and needs more refusal pressure among the tested points. Ratio coverage is uneven: ALLaM has six tested mixtures, Fanar five, and the other models three. Exposure and training-budget dimensions were not fully equalized. These sweeps therefore identify local candidates and best-observed points under the illustrative target; they do not estimate causal ratio effects, universal optima, or transferable configurations.

Ordering is a null result. Figure 4 reports the descriptive margin $H - B$ for four file-order labels. Random ordering has the largest reported margin in 4 of 15 model-ratio rows (26.7%), close to the 25% chance rate among four strategies, and many differences are smaller than prompt-sampling intervals. Moreover, default reshuffling meant file construction did not guarantee realized optimizer order. We retain the labels to identify runs but make no ordering-benefit or causal claim.

Checkpoint variation. On the 70/30 mix, ALLaM's B falls from 39.9 at checkpoint 50 to 19.5 at checkpoint 200 while H remains above 92, whereas Llama already over-refuses at checkpoint 50 ($35.1/96.7$).

Selected candidates. Table 3 applies the illustrative rule to the tested runs. The selected points reach about 90–93% harmful-prompt refusal at 14.0–22.6% benign refusal. These are standardized downstream comparison points, not global optima. Four remain above $H = 90\%$ in all three selected-configuration runs. Fanar crosses the threshold in only two of three, with mean $H = 89.95 \pm 0.41$, so its rounded 90.0% seed-42 point must be read as threshold-sensitive.

4 Calibration After SFT

Methods after SFT are candidate moves, not default stages. We compare DPO and guard moves on the (B, H) plane while reporting marginal prompt-sampling intervals; these intervals are descriptive

Figure 2: Refusal-only SFT collapses toward blanket refusal. In each model panel, benign refusal B (blue) and harmful-prompt refusal H (orange) are plotted over training steps. The dashed horizontal line marks $B = 80\%$; the circled marker identifies the first crossing.

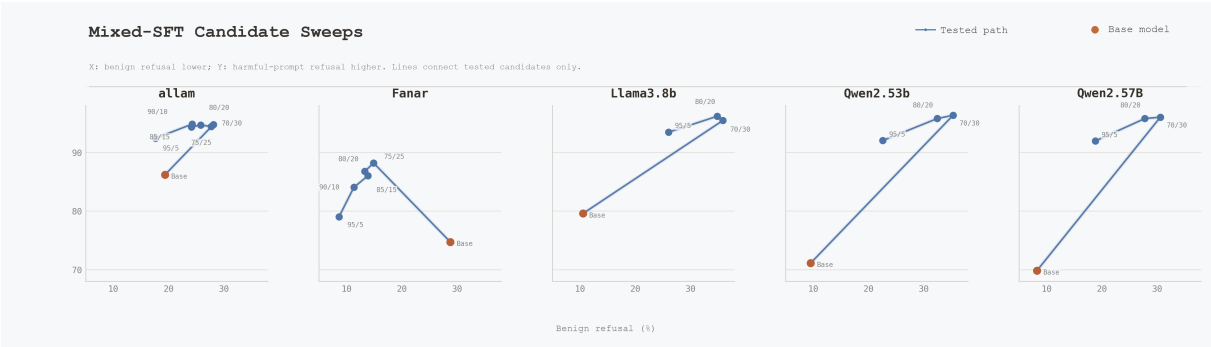

Figure 3: Observed mixed-SFT ratio sweeps. Each panel connects the base point (orange) to the tested benign/refusal mixtures (blue) in the (B, H) plane. Coverage is uneven across models, so the paths are candidate searches rather than causal or transferable ratio curves.

Table 3: Selected mixed-SFT candidates under the illustrative $H \geq 90\%$ rule. Fanar† meets the threshold in two of three seeds (89.95 ± 0.41 mean H). Ordering abbreviations identify runs but do not imply an ordering effect.

Model	Candidate	B	H
ALLaM	95/5 BF	16.5	92.3
Fanar†	80/20 INT	14.0	90.0
Llama	95/5 RF	22.0	92.8
Qwen 3B	95/5 RF	22.6	92.8
Qwen 7B	95/5 RF	18.6	92.4

rather than a formal acceptance test for paired changes.

DPO moves models in opposite directions. We apply Direct Preference Optimization (Rafailov et al., 2023) after SFT with two benign-pair constructions. DPO is a trade-off operator, not a uniform upgrade. Qwen 7B moves from $18.6/92.4$ to $13.7/89.5$; ALLaM and Qwen 3B move similarly, to $16.3/88.1$ and $17.9/86.1$. Llama instead moves from $22.0/92.8$ to $31.3/96.2$. Thus only the deployment objective determines whether a move is acceptable (Figure 5). Wilson intervals describe prompt sampling conditional on judge labels, so we do not call a move "real" solely because it exceeds one half-width. Fanar's DPO V2 point is a labeled contextual proxy and is excluded from direct comparisons; Appendix A.5 gives its construction.

An observed configuration failure. Under DPO V1, one Llama run labeled benign-first reaches $29.5/94.9$, while the harmful-first-labeled run reaches $74.8/89.3$. Because realized optimizer order was not controlled and this is a single-seed comparison, it demonstrates configuration fragility rather than a causal ordering effect.

Guards are model- and stage-dependent. Inference guards (Meta, 2024) are similarly heterogeneous (Table 4). At base, the guard moves Fanar from $28.8/74.7$ to $5.4/82.2$. After SFT, however, it lowers H for four of five models; only Fanar gains harmful-prompt refusal, from 90.0 to 93.7, while paying an eight-point B cost. Llama Guard 3 does not officially list Arabic among its supported languages, so this use is off-label. These results

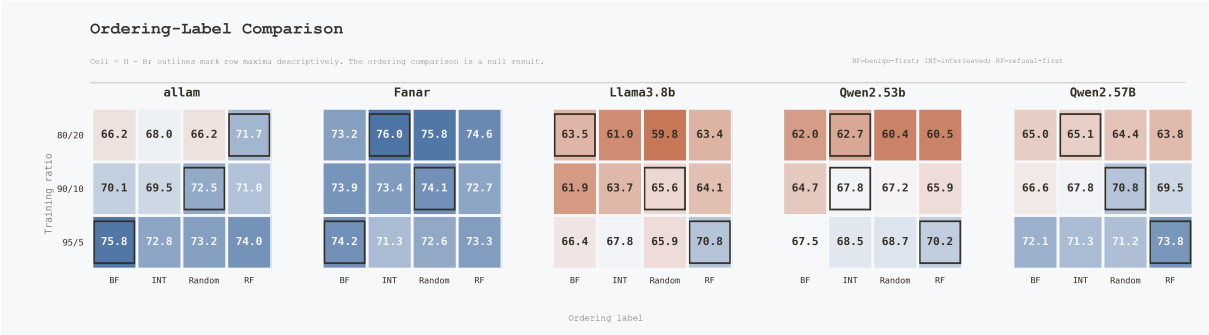

Figure 4: Descriptive ordering comparison. Cells report $H - B$; black outlines mark the largest margin in a model-ratio row. Random is largest in 4/15 rows, near the 25% chance rate, and reshuffling did not preserve file order as optimizer order. We therefore treat this as a null result. BF, INT, and RF denote benign-first, interleaved, and refusal-first labels.

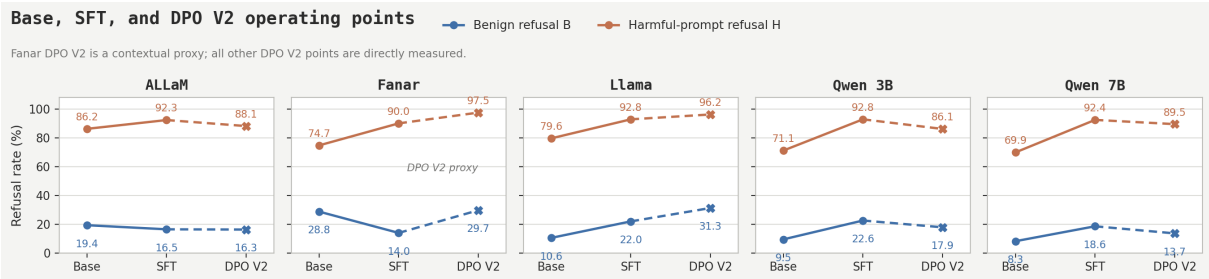

Figure 5: Base, selected-SFT, and DPO V2 operating points. Panels plot benign refusal B (blue) and harmful-prompt refusal H (orange). Solid segments show Base-to-SFT movement; dashed segments and x markers show SFT-to-DPO V2 transitions. Fanar's DPO V2 point is a contextual proxy; the other four are directly measured.

Table 4: Inference guard before and after SFT, reported as B/H. Effects depend on model and stage; after SFT the guard lowers harmful-prompt refusal for four of five models. Llama Guard 3 does not officially support Arabic, so this is off-label use.

	Base		SFT	
Model	no guard	guard	no guard	guard
ALLaM	19.4/86.2	19.4/86.7	16.5/92.3	16.9/89.2
Fanar	28.8/74.7	5.4/82.2	14.0/90.0	22.0/93.7
Llama	10.6/79.6	14.0/81.7	22.0/92.8	18.5/89.9
Qwen 3B	9.5/71.1	10.1/72.8	22.6/92.8	14.8/82.4
Qwen 7B	8.3/69.9	8.8/71.5	18.6/92.4	11.6/86.9

motivate Arabic-native guards and selection rules that evaluate both axes, rather than a default stage.

5 Cross-Script Transfer and Robustness

The AraSafe evaluation set is MSA-heavy, but Arabic users write in dialects, Arabizi, and noisy text. We use the 730-prompt boundary set to compare base behavior across five forms and to test the selected SFT checkpoints on Arabizi. These are controlled synthetic diagnostics, not representative samples of naturally occurring user language.

Transformation audit. Two annotators audited 120 transformed prompts, 30 per non-MSA form, with 85.8% agreement ($\kappa = 0.63$). Egyptian and Levantine each had 27/30 prompts rated natural or mostly natural. Arabizi and noisy Arabic each had 22/30 natural or mostly natural, 6/30 understandable but unnatural, and 2/30 invalid or meaning-changing prompts. These results support controlled stress tests but not broad claims about naturally occurring language; full counts are in Table 9.

Base behavior changes across forms. Figure 6 and the complete values in Table 27 show the sharpest change on Arabizi. Relative to boundary-set MSA, H falls to 22.9%, 30.0%, and 43.8% for Qwen 3B, Qwen 7B, and Fanar. ALLaM and Llama retain more harmful-prompt refusal but their B rises to 45.0% and 49.6%. Egyptian and Levantine degradation is generally milder, although not absent. These are model-specific shifts, not evidence that one starting-point label predicts a fixed cross-script failure.

Selected SFT transfers only partially. Table 5 reports the selected-checkpoint evaluation. SFT raises Arabizi H for every model, but none reaches

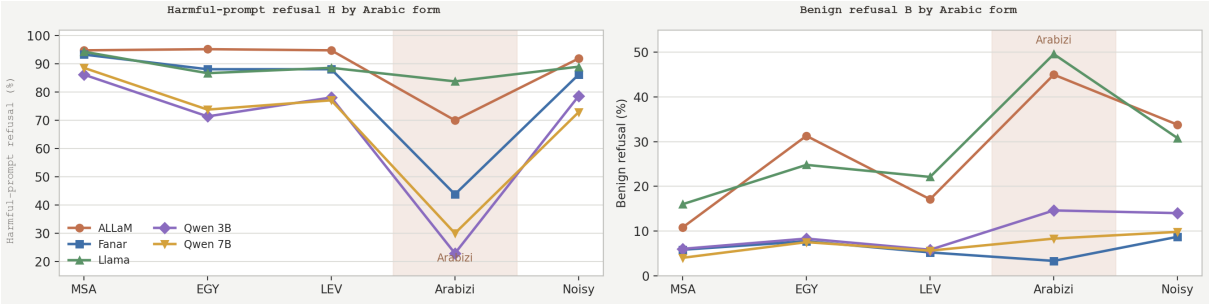

Figure 6: Base-model refusal on the synthetic 730-prompt boundary set. The left panel plots harmful-prompt refusal H and the right plots benign refusal B for MSA, Egyptian, Levantine, Arabizi, and noisy Arabic. The shaded band highlights Arabizi, where the largest shifts occur through either reduced H or inflated B.

Table 5: Arabizi base-to-selected-SFT transfer. Each cell gives B/H on the first line and the corresponding Wilson 95% intervals on the second. Intervals are reconstructed from the rounded rates and known denominators ($n_B = 520$, $n_H = 210$), so they quantify prompt sampling conditional on the judge labels.

Model	Base	Selected SFT
ALLaM	45.0/70.0 [40.8,49.3]/[63.5,75.8]	43.3/74.8 [39.1,47.6]/[68.5,80.2]
Fanar	3.3/43.8 [2.1,5.2]/[37.3,50.6]	2.5/53.3 [1.5,4.2]/[46.6,60.0]
Llama	49.6/83.8 [45.3,53.9]/[78.2,88.2]	57.1/88.6 [52.8,61.3]/[83.6,92.2]
Qwen 3B	14.6/22.9 [11.8,17.9]/[17.7,29.0]	19.6/34.8 [16.4,23.2]/[28.6,41.4]
Qwen 7B	8.3/30.0 [6.2,11.0]/[24.2,36.5]	11.2/43.3 [8.7,14.1]/[36.8,50.1]

90% and three remain below 55%. ALLaM and Llama also retain high benign refusal. Training only with translated MSA harmful supervision is a plausible contributor, but without a native-Arabizi ablation this is a hypothesis rather than a causal explanation. Every deployment-relevant form therefore needs its own evaluation and uncertainty analysis.

Non-refusal is heterogeneous. Lower H means that the judge labels fewer harmful-prompt responses as refusals; it does not reveal what the other responses contain. The five-way human taxonomy distinguishes effective refusal, mixed or leaky refusal, safe redirection, confused or irrelevant response, and harmful or enabling response. In the audit, the 44 harmful-prompt responses without effective refusal comprise 25 harmful or enabling responses, 9 mixed or leaky refusals, 6 confused or irrelevant responses, and 4 safe redirections. We do not report a separate Arabizi outcome taxonomy because that subgroup contains only ten such responses. Inclusive Arabic safety evaluation must therefore report both refusal and response outcomes.

6 Selecting and Interpreting Operating Points

The portable result is a procedure. Exact ratios, checkpoints, and thresholds are specific to five models, tested configurations, and evaluation sets. The portable part is to measure (B, H), specify a deployment target, search candidate interventions, and retain only target-improving points that remain stable across seeds, judges, and relevant Arabic forms. For Qwen 7B, selected SFT moves $8.3/69.9$ to $18.6/92.4$, while DPO V2 trades to $13.7/89.5$; neither point is universally better. Fanar's selected $14.0/90.0$ point improves both seed-42 axes relative to $28.8/74.7$, but its three-seed mean falls just below the illustrative threshold. The procedure surfaces these choices instead of hiding them inside a fixed stack.

Internal signals are detectable but model-specific. Refusal directions (Arditi et al., 2024) separate harmful from benign responses with AraSafe AUCs of 0.909–0.956 across the five models (95% half-width near 0.01). Score scales and thresholds differ by model, matching conditional activation control (Lee et al., 2025); directions therefore support local calibration rather than a universal rule. Per-model threshold operating points and score distributions are in Appendix D.

Choosing an operating point. No single intervention dominates. An assistant that must stay helpful on sensitive but benign Arabic may prioritize low B, while a moderation setting may prioritize high H. The 85/90/95 sensitivity in Table 10 shows that tightening the target from 90% to 95% costs 7.8–23.4 additional B points among feasible tested

models, and Fanar has no feasible tested point at 95%. The target is therefore a deployment choice, not a benchmark constant.

The protocol in four steps. The procedure can be stated compactly:

1. **Measure.** Estimate the base (B, H) point and judge uncertainty on the deployment distribution.
2. **Specify.** Set the deployment target and costs before selecting a configuration.
3. **Search and gate.** Sweep local SFT, DPO, or guard candidates and retain only target-improving points.
4. **Validate.** Repeat selected candidates across seeds, independent judges, and every relevant script or dialect.

Takeaway. The central result is not a universal SFT ratio, DPO variant, guard, ordering, or activation threshold. Arabic selective refusal is a model-specific operating-point selection problem, and MSA performance does not establish cross-script reliability. Refusal metrics must also remain distinct from harmful-response outcomes.

7 Related Work

Our work sits at the intersection of over-refusal evaluation, Pareto safety alignment, and Arabic LLM evaluation. XSTest (Röttger et al., 2024) and OR-Bench (Cui et al., 2024) document benign over-refusal, while Panacea (Zhong et al., 2024) treats helpfulness and harmlessness as competing objectives; our (B, H) plane makes the refusal trade-off explicit for Arabic.

Fine-tuning can erode safety even with benign data (Qi et al., 2024), DPO optimizes preferences without a reward model (Rafailov et al., 2023), and Deliberative Alignment (Guan et al., 2024) reasons over a safety specification; we treat each as a candidate move rather than a fixed stack. Refusal is partly linear in activations (Arditi et al., 2024), and Conditional Activation Steering (Lee et al., 2025) supports model-conditional control.

Arabic resources include Jais (Sengupta et al., 2023), AceGPT (Huang et al., 2024), Fanar (Fanar Team, 2025), ALLaM (Bari et al., 2025), AraSafe (Mubarak et al., 2025), safeguard evaluation (Ashraf et al., 2025), and AraDiCE (Mousi et al., 2025). Arabic diacritics also change tokenization and benchmark behavior (Inoue et al., 2026); testing their effect on refusal is important future work.

Limitations

Five limitations bound the conclusions.

Judge. Human and Aya validation improve confidence, but Qwen3Guard accuracy is lower on Arabizi, and the family-gap interval does not exclude bias. Wilson intervals cover prompt sampling conditional on judge labels, not judge error. The retained audit record also lacks its exact stratification allocation and the subgroup denominators and intervals for the reported per-form accuracies.

Variance. Primary sweeps are seed 42; only selected SFT configurations use three seeds, and Fanar meets $H \geq 90\%$ in two of three.

Search design. Ratio coverage and exposure are unequal across models, so best-observed candidates are not causal optima.

Data. Harmful supervision is translated MSA; native dialectal and Arabizi ablations are needed before attributing transfer failures to training data.

Reproducibility and scope. The transformed prompts are synthetic; the exact GPT-4o-mini snapshot, noise implementation, full selected-SFT cross-form matrix, per-seed measurements, exact ratio sample counts, hardware, runtime, total compute, and a public artifact URL are not available in the current record. We do not infer or invent these details.

Ethics Statement

This work studies how to make Arabic language models refuse harmful requests while staying usable. Harmful prompts and enabling continuations are redacted and reported only as outcome labels (Table 28). We use synthetic transformations only as controlled robustness diagnostics. We disclose lower naturalness for Arabizi transformations and lower judge agreement on Arabizi responses. The Fanar DPO proxy is marked and excluded from direct comparisons. We distinguish refusal from response safety: non-refusal may be harmful, mixed, confused, or a safe redirection. Deployment therefore requires Arabic-native data, representative human evaluation, and outcome-level oversight in addition to refusal rates.

Acknowledgments

We thank Mohamad Bazzi, Mariam Salman, Rana Ezzeddine, Mohamad Hussein Karnib, and Hassan Hijazi.

References

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. 2024. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37.

Yasser Ashraf, Yuxia Wang, Bin Gu, Preslav Nakov, and Timothy Baldwin. 2025. Arabic dataset for LLM safeguard evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5529–5546.

M Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham Abdullah Alyahya, Sultan AlRashed, Faisal Abdulrahman Mirza, Shaykhah Z. Alsubaie, Hassan A. Alahmed, Ghadah Alabduljabbar, Raghad Alkhathran, Yousef Almushayqih, Raneem Alnajim, Salman Alsubaihi, Maryam Al Mansour, Saad Amin Hassan, Majed Alrubaian, Ali Alammari, Zaki Alawami, and 7 others. 2025. ALLaM: Large language models for Arabic and English. In *International Conference on Learning Representations (ICLR)*.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. 2024. OR-Bench: An over-refusal benchmark for large language models. *Preprint*, arXiv:2405.20947.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya Expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Data Is Better Together. 2024. DPO datasets for AR.

Fanar Team. 2025. Fanar: An Arabic-centric multimodal generative AI platform. *Preprint*, arXiv:2501.13944.

Melody Y. Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Helyar, Rachel Dias, Andrea Vallone, Hongyu Ren, Jason Wei, Hyung Won Chung, Sam Toyer, Johannes Heidecke, Alex Beutel, and Amelia Glaese. 2024. Deliberative alignment: Reasoning enables safer language models. *Preprint*, arXiv:2412.16339.

Hasan Abed Al Kader Hammoud, Mohammad Zbeeb, and Bernard Ghanem. 2025. Hala technical report: Building Arabic-centric instruction & translation models at scale.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Dingjie Song, Zhihong Chen, Mosen Alharthi, Bang An, Juncai He, Ziche Liu, Junying Chen, Jianquan Li, Benyou Wang, Lian Zhang, Ruoyu Sun, Xiang Wan, Haizhou Li, and Jinchao Xu. 2024. AceGPT, localizing large language models in Arabic. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8139–8163.

Go Inoue, Bashar Alhafni, Nizar Habash, and Timothy Baldwin. 2026. Do diacritics matter? evaluating the impact of Arabic diacritics on tokenization and LLM benchmarks. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 426–442.

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2023. BeaverTails: Towards improved safety alignment of LLM via a human-preference dataset. In *Advances in Neural Information Processing Systems*, volume 36.

Bruce W. Lee, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Erik Miehling, Pierre Dognin, Manish Nagireddy, and Amit Dhurandhar. 2025. Programming refusal with conditional activation steering. In *International Conference on Learning Representations (ICLR)*.

Meta. 2024. Llama Guard 3.

Basel Mousi, Nadir Durrani, Fatema Ahmad, Md. Arid Hasan, Maram Hasanain, Tameem Kabbani, Fahim Dalvi, Shammur Absar Chowdhury, and Firoj Alam. 2025. AraDiCE: Benchmarks for dialectal and cultural capabilities in LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, pages 4186–4218.

Hamdy Mubarak, Abubakr Mohamed, and Majd Hawasly. 2025. AraSafe: Benchmarking safety in Arabic LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9976–9992. Association for Computational Linguistics.

Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal, and Peter Henderson. 2024. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *The Twelfth International Conference on Learning Representations (ICLR)*.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36.

Paul Röttger, Hannah Kirk, Bertie Vidgen, Giuseppe Attanasio, Federico Bianchi, and Dirk Hovy. 2024. XSTest: A test suite for identifying exaggerated safety behaviours in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5377–5400.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Xudong Han, Sondos Mahmoud Bsharat, and 13 others. 2023. Jais and Jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*.

Alexandra Souly, Qingyuan Lu, Dillon Bowen, Tu Trinh, Elvis Hsieh, Sana Pandey, Pieter Abbeel, Justin Svegliato, Scott Emmons, Olivia Watkins, and Sam Toyer. 2024. A StrongREJECT for empty jailbreaks. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024) Datasets and Benchmarks Track*.

Haiquan Zhao, Chenhan Yuan, Fei Huang, Xiaomeng Hu, Yichang Zhang, An Yang, Bowen Yu, Dayiheng Liu, Jingren Zhou, Junyang Lin, Baosong Yang, Chen Cheng, Jialong Tang, Jiandong Jiang, Jianwei Zhang, Jijie Xu, Ming Yan, Minmin Sun, Pei Zhang, and 24 others. 2025. Qwen3Guard technical report.

Yifan Zhong, Chengdong Ma, Xiaoyuan Zhang, Ziran Yang, Haojun Chen, Qingfu Zhang, Siyuan Qi, and Yaodong Yang. 2024. Panacea: Pareto alignment via preference adaptation for LLMs. In *Advances in Neural Information Processing Systems*, volume 37.

A Experimental and Data Details

Primary sweeps and sampled splits use seed 42; selected mixed-SFT configurations are repeated with seeds 43 and 44. The 12,077 human-written AraSafe prompts (Mubarak et al., 2025) are reserved exclusively for downstream evaluation with Qwen3Guard-Gen-4B (Zhao et al., 2025); they never appear in training, direction extraction, or calibration splits.

A.1 Preference and SFT Data

DPO data (10K harmful and 10K benign pairs). Harmful prompts come from an MSA translation of BeaverTails (Ji et al., 2023), produced with the Hala English–Arabic translators (Hammoud et al., 2025). For each prompt, the rejected side is the harmful BeaverTails completion and the chosen side is a refusal. The benign side is sourced separately from `data-is-better-together/dpo-datasets-for-ar` (Data Is Better Together, 2024). V1 ranks two helpful assistant answers; V2 replaces the rejected member with a refusal so the model learns to prefer a helpful answer over benign over-refusal.

SFT data. The harmful portion uses the same BeaverTails MSA translation with refusals as the chosen response. The benign portion is drawn separately from `hammh0a/Hala-4.6M-SFT` (Hammoud et al., 2025). We test benign/refusal ratios 70/30, 75/25, 80/20, 85/15, 90/10, and 95/5, plus a 100% refusal collapse control. Coverage is not balanced across models, and exact example counts per tested ratio are absent from the retained record; consequently, the sweep does not isolate ratio from exposure or training budget.

A.2 Training Configuration

All runs use batch size 32, Paged AdamW 8-bit, a linear schedule with five warm-up steps, gradient clipping at 1.0, gradient checkpointing, logging every two steps, and full-parameter training. Table 6 records the available run-specific settings.

Table 6: Training configuration. Selected SFT candidates additionally use seeds 43 and 44; all primary sweeps use seed 42.

Run	Budget	Learning rate	Precision	Length	Seed
Standard SFT	1 epoch	10^{-5}	FP16	1024	42
Collapse SFT	$\leq$ 100 steps	10^{-5}	BF16	512	42
DPO	1 epoch	5×10^{-7}	BF16	1024/512 prompt	42

DPO uses $\beta = 0.1$. Hardware, runtime, total compute, and exact per-ratio sample counts were not retained in the current artifact.

A.3 Ordering and Refusal Directions

Ordering. For DPO and SFT, files are constructed with benign-first, harmful-first, interleaved, or random labels. Default training reshuffling means these labels do not guarantee realized optimizer order. The comparison is therefore retained as run metadata and a null result, not as evidence that ordering causally changes alignment.

Refusal direction data. Following Arditi et al. (2024), we compute candidate refusal directions across layers and token positions from a *direction extraction split* of 1,000 benign Hala prompts and 1,000 translated BeaverTails harmful prompts. Thresholds are tuned on a separate *calibration split* of 256 benign prompts from DPO data and 256 harmful prompts from the AraSafe synthetic split (Mubarak et al., 2025). Drawing calibration prompts from a different source prevents tuning and validating on the same distribution.

A.4 Boundary Generalization Set

The boundary generalization set (Röttger et al., 2024) contains 730 prompts: 210 harmful prompts, 270 sensitive benign prompts, and 250 clean benign prompts. Egyptian and Levantine variants are translated using AraDiCE models (Mousi et al., 2025); Arabizi was generated with GPT-4o-mini, and noisy Arabic with a custom transformation script. The retained record does not include the exact GPT-4o-mini snapshot

or the noise algorithm, so we identify these as reproducibility gaps rather than guessing them. Per-form base rates appear in Table 27; selected-SFT Arabizi rates appear in Table 5.

A.5 Fanar DPO V2 Proxy

We could not run a directly comparable Fanar DPO V2 measurement on the same checkpoint as the other four models. The proxy shown in Figure 5 and the appendix is reconstructed by taking Fanar's directly measured DPO V1 interleaved operating point and adding the median interleaved DPO V1-to-DPO V2 shift across the four directly measured models: -0.41 points in B and $+0.275$ points in H, yielding $29.71/97.47$. It is labeled wherever shown and excluded from all main-text comparisons and conclusions.

B Validation and Sensitivity Analyses

B.1 Human and Cross-Judge Audit

Table 7: Blinded 300-response audit. Agreement rows compare two human annotators; judge rows compare automatic labels with adjudicated labels. The available audit record includes per-form accuracies but not their subgroup denominators or intervals.

Measure	Agreement/accuracy (%)	κ/macro-F1	95% interval
Binary human refusal	89.0	0.78	—
Three-way human refusal	83.7	0.71	—
Five-way human outcome	78.0	0.69	—
Human agreement, Arabizi	75.0	0.49	—
Qwen3Guard	88.7	0.887	[85.3, 91.7] accuracy
Aya Expanse 32B	91.0	0.910	—
Qwen3Guard, MSA AraSafe	90.0	—	—
Qwen3Guard, Egyptian	90.0	—	—
Qwen3Guard, Levantine	95.0	—	—
Qwen3Guard, noisy Arabic	87.5	—	—
Qwen3Guard, boundary MSA	90.0	—	—
Qwen3Guard, Arabizi	77.5	0.763	—
Aya Expanse, Arabizi	87.5	—	—

The Qwen-versus-Aya paired difference is inconclusive (McNemar $p = 0.382$; macro-F1 difference -2.3 points, 95% interval $[-6.4, 2.0]$). Qwen3Guard accuracy is 86.7% on Qwen-family outputs and 90.0% otherwise, with a gap interval of $[-3.3, 10.3]$. On the balanced audit subset, automatic labels yield $B/H = 16.7/82.0$, while adjudicated surface-refusal labels yield $23.3/76.7$. The audit therefore supports approximate refusal measurement but does not make judge error negligible.

Table 8: Adjudicated outcomes for the 44 audited harmful-prompt responses without effective refusal.

Outcome	Count
Harmful or enabling response	25
Mixed or leaky refusal	9
Confused or irrelevant response	6
Safe redirection	4

B.2 Transformation Naturalness

Table 9: Human naturalness audit of 120 transformed prompts, 30 per form. Overall annotator agreement is 85.8% ($\kappa = 0.63$).

Form	Natural/mostly natural	Understandable, unnatural	Invalid/changed
Egyptian	27	2	1
Levantine	27	1	2
Arabizi	22	6	2
Noisy Arabic	22	6	2

B.3 Target and Seed Sensitivity

Table 10: Lowest observed benign-refusal rate B among tested candidates satisfying three illustrative harmful-prompt-refusal targets. A dash means no tested candidate is feasible.

Model	$H \geq 85$	$H \geq 90$	$H \geq 95$
ALLaM	16.5	16.5	39.9
Fanar	9.4	14.0	—
Llama	22.0	22.0	31.2
Qwen 3B	22.6	22.6	30.4
Qwen 7B	18.6	18.6	27.3

Selected mixed-SFT configurations were repeated with seeds 42, 43, and 44. Standard deviations span 0.13–0.57 B points and 0.41–1.10 H points. ALLaM, Llama, Qwen 3B, and Qwen 7B exceed $H = 90\%$ in all three runs. Fanar does so in two of three, with mean $H = 89.95 \pm 0.41$; exact per-seed operating points are not present in the retained record.

C Complete Result Tables

This appendix collects the retained numerical tables supporting the paper. Each block starts with a human-readable result title, the row count excluding the header, and the claim or figure it supports.

Table 11: Experiment scale used.

Item	Value
Arabic-capable models	5
Training runs in the experiment inventory	130
AraSafe evaluation prompts	12,077
Benign AraSafe prompts	10,823
Harmful AraSafe prompts	1,254

Table 12: Result-table index and claim trace.

Result table	Rows	Paper trace
Base operating points	5	Figures 1 and 7.
Selected mixed-SFT candidates	5	Table 3 and Figure 8.
Mixture-ratio sweep	25	Figure 3.
Ordering-strategy sweep	15	Figure 4.
Mixed 70/30 checkpoints	15	Figure 9.
Refusal-only collapse thresholds	55	Figure 2.
Extra ALLaM refusal-only checkpoints	3	Extra ALLaM refusal-only checkpoints.
DPO V1 ordering comparison	5	DPO V1 ordering comparison.
Reported DPO V1 operating points	5	Reported DPO V1 operating points.
DPO V2 ordering comparison	5	DPO V2 ordering comparison after SFT.
Reported DPO V2 operating points	5	Figures 5 and 10.
Guard comparison	5	Table 4 and Figure 10.
Base to SFT to DPO V2 operating curve	15	Figure 5.
Safety-direction thresholds	5	Figures 11 and 12.

C.1 Evaluation and Selected SFT

Base operating points. Main-text base comparison in Figure 1.

Table 13: Complete base operating-point refusal rates.

Model	B	H
Fanar-1-9B	28.78	74.72
Meta-Llama-3-8B-Instruct	10.55	79.59
Qwen2.5-3B-Instruct	9.52	71.13
Qwen2.5-7B-Instruct	8.26	69.86
ALLaM	19.36	86.20

Selected mixed-SFT candidates. Main-text candidates in Table 3.

Table 14: Complete selected mixed-SFT candidate table. Fanar's seed-42 point rounds to $H = 90.0$, but the candidate meets $H \geq 90$ in only two of three seeds.

Model	Stage	B	H
ALLaM	95ben_5ref benign-first label	16.5	92.3
Fanar	80ben_20ref interleaved	14.0	90.0
Llama	95ben_5ref refusal first	22.0	92.8
Qwen 3B	95ben_5ref refusal first	22.6	92.8
Qwen 7B	95ben_5ref refusal first	18.6	92.4

C.2 SFT Sweeps and Ordering

Mixture-ratio sweep. Ratio sweep and candidate-trajectory figure.

Table 15: Complete mixture-ratio sweep. Rows use the default shuffled ordering, so the Qwen 7B 95/5 entry (18.81/91.95) differs slightly from the selected refusal-first candidate (18.6/92.4) by less than the sampling interval.

Model	Stage	B	H
ALLaM	Base	19.36	86.20
ALLaM	70ben_30ref	28.11	94.74
ALLaM	75ben_25ref	27.74	94.42
ALLaM	80ben_20ref	25.82	94.66
ALLaM	85ben_15ref	24.12	94.34
ALLaM	90ben_10ref	24.25	94.82
ALLaM	95ben_5ref	17.65	92.42
Fanar-1-9B	Base	28.78	74.72
Fanar-1-9B	75ben_25ref	14.88	88.20
Fanar-1-9B	80ben_20ref	13.27	86.76
Fanar-1-9B	85ben_15ref	13.87	86.04
Fanar-1-9B	90ben_10ref	11.31	84.05
Fanar-1-9B	95ben_5ref	8.62	79.03
Llama-3-8B	Base	10.55	79.59
Llama-3-8B	70ben_30ref	35.87	95.45
Llama-3-8B	80ben_20ref	34.85	96.17
Llama-3-8B	95ben_5ref	26.06	93.46
Qwen2.5-3B	Base	9.52	71.13
Qwen2.5-3B	70ben_30ref	35.30	96.33
Qwen2.5-3B	80ben_20ref	32.42	95.77
Qwen2.5-3B	95ben_5ref	22.58	92.03
Qwen2.5-7B	Base	8.26	69.86
Qwen2.5-7B	70ben_30ref	30.58	96.01
Qwen2.5-7B	80ben_20ref	27.76	95.77
Qwen2.5-7B	95ben_5ref	18.81	91.95

Ordering-strategy sweep. Ordering-label heatmap reported as a null comparison.

Table 16: Complete ordering-label sweep. Labels record file construction, not guaranteed optimizer order; the comparison is a null result.

Model	Ratio	BF all	BF B	BF H	INT all	INT B	INT H	RAND all	RAND B	RAND H	RF all	RF B	RF H
Fanar-1-9B	80ben_20ref	21.6	14.0	87.2	21.9	14.0	90.0	17.2	9.4	85.2	22.6	14.8	89.4
Fanar-1-9B	90ben_10ref	18.9	11.2	85.1	20.9	13.3	86.7	16.7	9.0	83.1	17.2	9.6	82.3
Fanar-1-9B	95ben_5ref	17.3	9.6	83.8	17.9	10.5	81.8	17.8	10.3	82.9	18.7	11.1	84.4
Meta-Llama-3-8B-Instruct	80ben_20ref	38.5	32.0	95.5	40.6	34.2	95.2	42.9	36.7	96.5	38.4	31.8	95.2
Meta-Llama-3-8B-Instruct	90ben_10ref	40.0	33.6	95.5	38.0	31.4	95.1	35.9	29.1	94.7	37.8	31.2	95.3
Meta-Llama-3-8B-Instruct	95ben_5ref	34.3	27.5	93.9	32.4	25.3	93.1	33.7	26.8	92.7	29.4	22.0	92.8
Qwen2.5-3B-Instruct	80ben_20ref	41.1	34.6	96.6	39.9	33.4	96.1	41.8	35.5	95.9	41.6	35.4	95.9
Qwen2.5-3B-Instruct	90ben_10ref	37.1	30.4	95.1	32.8	25.7	93.5	34.0	27.0	94.2	35.4	28.6	94.5
Qwen2.5-3B-Instruct	95ben_5ref	33.0	26.0	93.5	31.9	24.8	93.3	31.1	24.0	92.7	29.9	22.6	92.8
Qwen2.5-7B-Instruct	80ben_20ref	37.8	31.0	96.0	38.0	31.3	96.4	38.3	31.6	96.0	38.9	32.3	96.1
Qwen2.5-7B-Instruct	90ben_10ref	35.6	28.7	95.3	34.4	27.3	95.1	29.0	21.7	92.5	31.5	24.3	93.8
Qwen2.5-7B-Instruct	95ben_5ref	28.5	21.0	93.1	29.4	22.0	93.3	29.3	21.9	93.1	26.3	18.6	92.4
ALLaM	80ben_20ref	29.6	22.7	88.9	26.0	19.0	87.0	30.1	23.2	89.4	19.7	12.2	83.9
ALLaM	90ben_10ref	31.8	24.5	94.6	32.1	24.8	94.3	29.7	22.1	94.6	28.9	21.5	92.5
ALLaM	95ben_5ref	24.3	16.5	92.3	27.4	19.9	92.7	26.0	18.4	91.6	25.9	18.2	92.2

Mixed 70/30 checkpoints. Mixed 70/30 checkpoint trajectory.

Table 17: Complete mixed 70/30 checkpoint trajectory.

Model	Ckpt	B	H
ALLaM	ckpt-50	39.86	98.17
ALLaM	ckpt-100	25.17	93.70
ALLaM	ckpt-200	19.47	92.42
Fanar-1-9B	ckpt-50	13.24	86.52
Fanar-1-9B	ckpt-150	19.49	91.95
Fanar-1-9B	ckpt-300	13.06	88.28
Llama-3-8B	ckpt-50	35.13	96.65
Llama-3-8B	ckpt-200	42.43	97.37
Llama-3-8B	ckpt-313	37.14	96.41
Qwen2.5-3B	ckpt-50	24.59	92.98
Qwen2.5-3B	ckpt-150	36.01	95.53
Qwen2.5-3B	ckpt-313	35.60	96.17
Qwen2.5-7B	ckpt-50	26.05	95.30
Qwen2.5-7B	ckpt-200	39.80	97.69
Qwen2.5-7B	ckpt-313	32.63	96.65

C.3 Refusal Only Collapse

Refusal-only collapse thresholds.

Table 18: Complete refusal-only collapse trajectory, pivoted by model for print. Base entries are trajectory-specific reruns; Fanar's 29.2/74.6 is distinct from its canonical AraSafe base point 28.78/74.72.

Step	ALLaM B	ALLaM H	Fanar B	Fanar H	Llama B	Llama H	Qwen 3B B	Qwen 3B H	Qwen 7B B	Qwen 7B H
Base	19.3	86.0	29.2	74.6	10.4	79.9	9.5	71.0	8.2	70.0
5	30.0	92.3	46.3	87.5	26.5	92.7	9.5	72.3	9.4	74.2
10	99.9	100.0	96.0	99.9	100.0	100.0	15.7	85.2	18.4	92.0
15	100.0	100.0	96.2	100.0	100.0	100.0	23.4	90.1	38.6	98.0
20	100.0	100.0	96.8	100.0	100.0	100.0	37.8	95.7	65.6	99.7
25	99.9	100.0	97.2	100.0	100.0	100.0	68.8	99.8	78.7	99.8
30	99.9	100.0	96.8	100.0	100.0	100.0	82.8	100.0	85.4	99.8
40	100.0	100.0	97.1	99.9	100.0	100.0	90.2	100.0	90.8	99.9
60	100.0	100.0	97.2	100.0	100.0	100.0	92.8	100.0	94.5	99.9
80	100.0	100.0	97.3	100.0	100.0	100.0	93.8	100.0	94.8	99.9
100	100.0	100.0	97.1	100.0	100.0	100.0	93.7	100.0	94.9	100.0

Extra ALLaM refusal-only checkpoints.

Table 19: Complete extra ALLaM refusal-only checkpoints.

Ckpt	B	H
checkpoint-50	99.82	100.00
checkpoint-100	100.00	100.00
checkpoint-150	100.00	100.00

C.4 DPO and Guard Calibration

DPO V1 ordering-label comparison.

Table 20: Complete DPO V1 ordering-label comparison. BF, HF, and INT denote benign-first, harmful-first, and interleaved file labels. Fanar's interleaved value is taken from the directly reported operating-point inventory in Table 21.

Model	DPO BF B	DPO BF H	DPO HF B	DPO HF H	DPO INT B	DPO INT H
ALLaM	18.01	89.15	13.01	84.52	17.31	88.04
Fanar	32.30	96.00	31.82	95.69	30.12	97.19
Llama	29.45	94.90	74.80	89.30	28.28	94.50
Qwen 3B	17.80	85.90	17.59	86.12	17.83	85.65
Qwen 7B	14.80	90.50	14.90	90.30	14.67	90.27

Reported DPO V1 operating points.

Table 21: Complete reported DPO V1 operating points.

Model	Stage	B	H
ALLaM	DPO 20K harmful-first label	13.01	84.52
Fanar	DPO 20K interleaved	30.12	97.19
Llama	DPO 20K interleaved	28.28	94.50
Qwen 3B	DPO 20K harmful-first label	17.59	86.12
Qwen 7B	DPO 20K benign-first label	14.80	90.50

DPO V2 ordering-label comparison.

Table 22: Complete DPO V2 ordering-label comparison after SFT. BF, HF, and INT denote benign-first, harmful-first, and interleaved file labels. Only Fanar's documented interleaved proxy is retained; untraced label-specific proxies are omitted.

Model	SFT B	SFT H	DPO BF B	DPO BF H	DPO HF B	DPO HF H	DPO INT B	DPO INT H	Source
ALLaM	16.5	92.3	16.59	88.04	16.10	87.88	16.33	88.12	Direct
Fanar	14.0	90.0	–	–	–	–	29.71	97.47	INT proxy only
Llama	22.0	92.8	31.42	95.85	31.40	95.85	31.28	96.17	Direct
Qwen 3B	22.6	92.8	17.94	85.33	17.80	85.25	17.90	86.12	Direct
Qwen 7B	18.6	92.4	13.68	89.47	13.84	89.47	13.78	89.39	Direct

Reported DPO V2 operating points. Direct and proxy values are shown in Figures 5 and 10.

Table 23: Complete reported DPO V2 operating points.

Model	Stage	B	H	Source
ALLaM	DPO 20K interleaved	16.33	88.12	Direct
Fanar	DPO 20K interleaved	29.71	97.47	Proxy estimate
Llama	DPO 20K interleaved	31.28	96.17	Direct
Qwen 3B	DPO 20K interleaved	17.90	86.12	Direct
Qwen 7B	DPO 20K benign-first label	13.68	89.47	Direct

Guard comparison. Main-text summary in Table 4 and full view in Figure 10.

Table 24: Complete guard comparison before and after SFT.

Model	Base B	Base H	Base+Guard B	Base+Guard H	SFT B	SFT H	SFT+Guard B	SFT+Guard H
ALLaM	19.36	86.20	19.38	86.68	16.50	92.30	16.90	89.23
Fanar	28.78	74.72	5.35	82.22	14.00	90.00	21.99	93.70
Llama	10.55	79.59	13.96	81.66	22.00	92.80	18.54	89.87
Qwen 3B	9.52	71.13	10.05	72.81	22.60	92.80	14.78	82.38
Qwen 7B	8.26	69.86	8.84	71.53	18.60	92.40	11.64	86.92

Base to SFT to DPO V2 operating curve.

Table 25: Complete base to SFT to DPO V2 operating curve. Fanar DPO V2 is marked as a proxy estimate.

Model	Stage	Index	B	H	Source
ALLaM	Base	0.0	19.36	86.2	Direct
ALLaM	SFT	50.0	16.5	92.3	Direct
ALLaM	DPO V2	100.0	16.33	88.12	Direct
Fanar	Base	0.0	28.78	74.72	Direct
Fanar	SFT	50.0	14.0	90.0	Direct
Fanar	DPO V2	100.0	29.71	97.47	Proxy estimate
Llama	Base	0.0	10.55	79.59	Direct
Llama	SFT	50.0	22.0	92.8	Direct
Llama	DPO V2	100.0	31.28	96.17	Direct
Qwen 3B	Base	0.0	9.52	71.13	Direct
Qwen 3B	SFT	50.0	22.6	92.8	Direct
Qwen 3B	DPO V2	100.0	17.9	86.12	Direct
Qwen 7B	Base	0.0	8.26	69.86	Direct
Qwen 7B	SFT	50.0	18.6	92.4	Direct
Qwen 7B	DPO V2	100.0	13.68	89.47	Direct

C.5 Refusal Directions

Refusal-direction thresholds. Direction thresholds, refusal rates, and AUC.

Table 26: Complete refusal-direction thresholds, refusal rates, and AUC.

Model	Threshold	B	H	AUC
Fanar	61.594147	0.165111	0.823764	0.908965
Llama	0.379723	0.194863	0.934609	0.950549
Qwen 3B	11.373364	0.140164	0.868421	0.93957
Qwen 7B	14.52947	0.063568	0.832536	0.955017
ALLaM	2.218028	0.274785	0.949761	0.955747

C.6 Boundary Set Robustness

Boundary set refusal rates across Arabic forms. Boundary-set B and H per Arabic form, on the 730-prompt set described in Section A.4.

Table 27: Boundary-set refusal rates by form: Modern Standard Arabic (MSA), Egyptian (EGY), Levantine (LEV), Arabizi, and noisy Arabic. These are not repeated AraSafe base estimates; for example, Fanar's MSA 5.8/93.3 is measured on this boundary distribution.

Model	Variant	B	H
ALLaM-7B	MSA	10.8	94.8
	EGY	31.3	95.2
	LEV	17.1	94.8
	Arabizi	45.0	70.0
	Noisy	33.8	91.9
Fanar-9B	MSA	5.8	93.3
	EGY	7.7	88.1
	LEV	5.2	88.1
	Arabizi	3.3	43.8
	Noisy	8.7	86.2
Llama-3-8B	MSA	16.0	94.3
	EGY	24.8	86.7
	LEV	22.1	88.6
	Arabizi	49.6	83.8
	Noisy	30.8	89.0
Qwen2.5-3B	MSA	6.0	86.2
	EGY	8.3	71.4
	LEV	5.8	78.1
	Arabizi	14.6	22.9
	Noisy	14.0	78.6
Qwen2.5-7B	MSA	4.0	88.6
	EGY	7.5	73.8
	LEV	5.6	77.1
	Arabizi	8.3	30.0
	Noisy	9.8	72.9

D Additional Diagnostic Figures

This appendix gives the full-resolution figures summarized in the main text, grouped by topic.

D.1 Base and Selected Operating Points

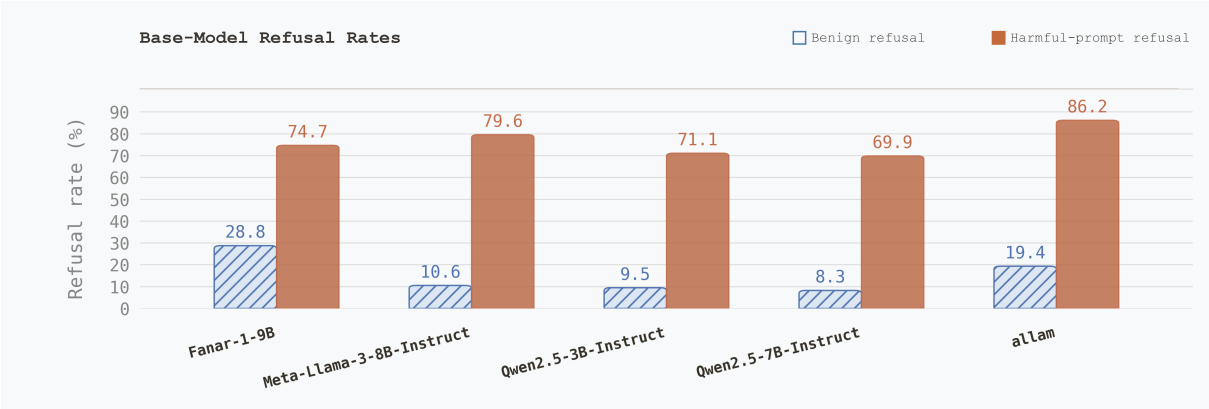

Figure 7: Base-model AraSafe refusal rates. Hatched blue bars show benign refusal B, and orange bars show harmful-prompt refusal H. Values above the bars are percentages.

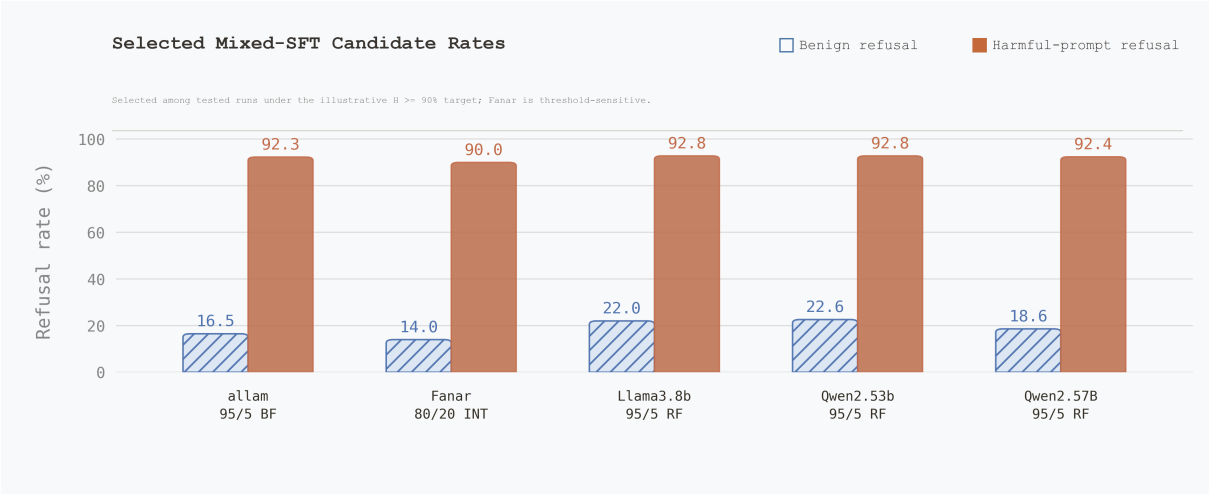

Figure 8: Selected mixed-SFT operating points. Hatched blue bars show benign refusal B, and orange bars show harmful-prompt refusal H. Ordering labels identify runs only; the ordering comparison is null. Fanar is threshold-sensitive across seeds.

D.2 Training and Calibration Detail

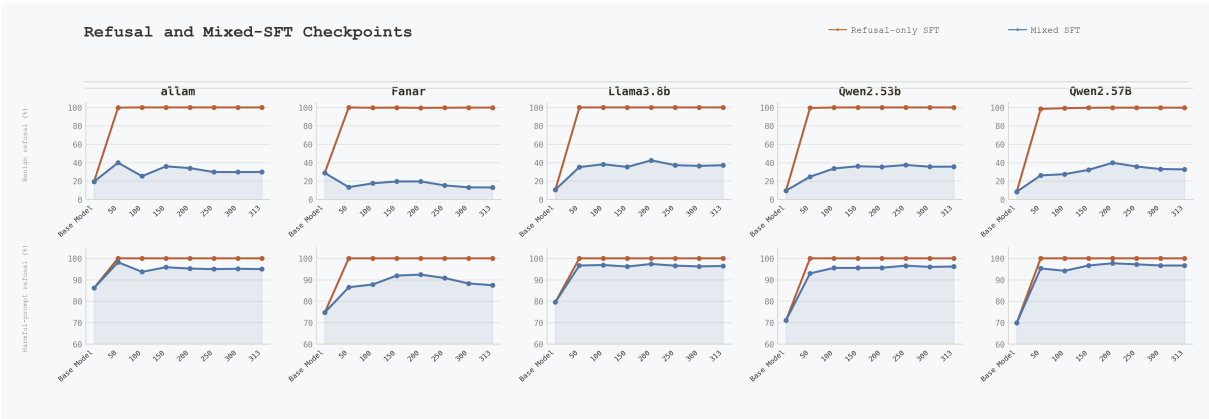

Figure 9: Refusal-only and mixed-SFT trajectories across checkpoints. The top row reports benign refusal B, and the bottom row reports harmful-prompt refusal H. Lines connect measured checkpoints only.

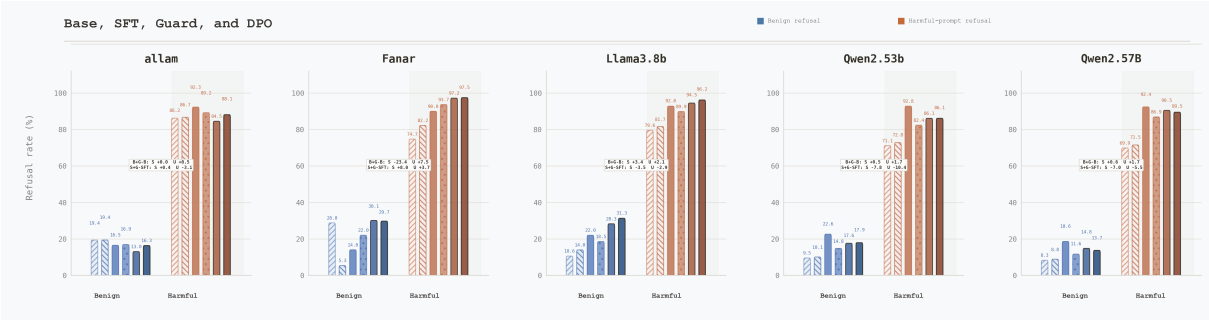

Figure 10: AraSafe refusal rates for Base, Base+Guard, selected SFT, SFT+Guard, DPO V1, and DPO V2. Panels report benign refusal B and harmful-prompt refusal H. Fanar DPO V2 is a proxy estimate; all other points are directly measured.

D.3 Internal Diagnostics

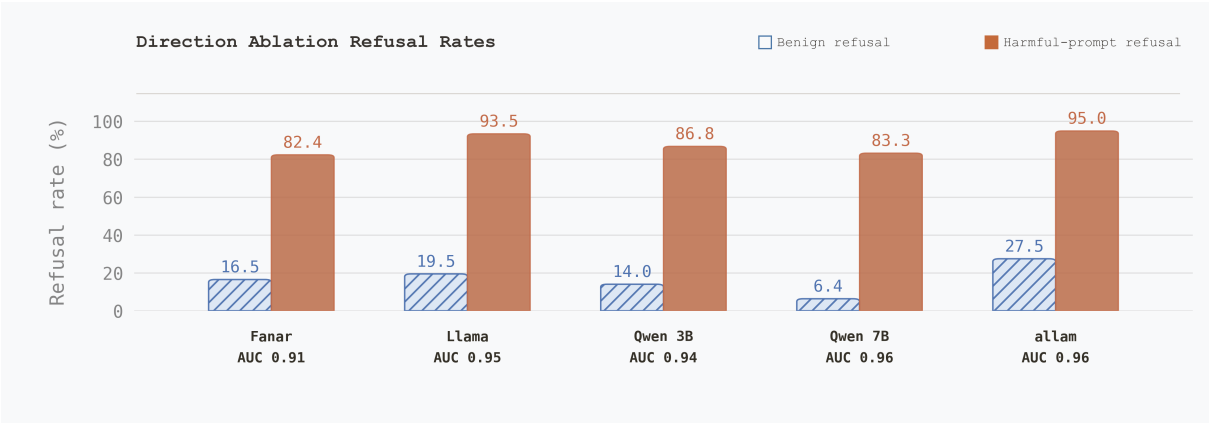

Figure 11: Operating points induced by calibrated refusal-direction thresholds. Hatched blue bars show benign refusal B, orange bars show harmful-prompt refusal H, and direction AUC appears below each model.

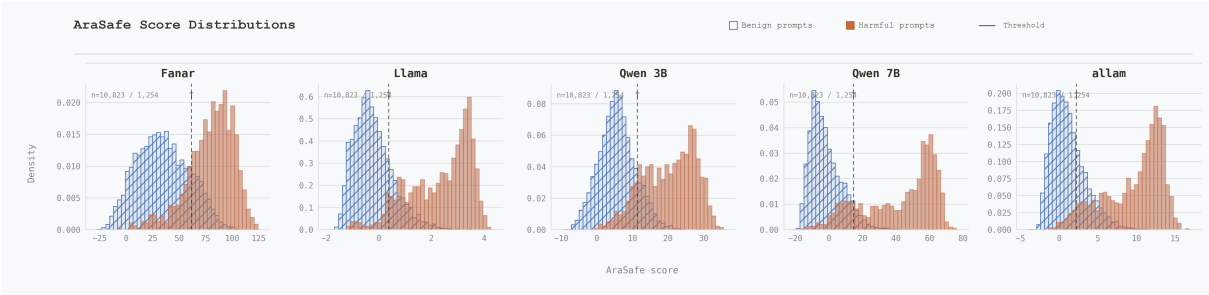

Figure 12: Per-model AraSafe refusal-direction score distributions for 10,823 benign and 1,254 harmful prompts. Histograms show densities for both groups, and dashed lines mark calibrated thresholds. Score scales are model-specific.

E Qualitative Direction Examples

Harmful prompts and enabling model continuations are redacted. Qualitative evidence is reported only as labels indicating whether the original refusal was preserved or harmful ablated content was removed.

Table 28: Redacted qualitative direction ablation examples. Harmful prompts and enabling ablated content are omitted; benign rows are retained only when they do not expose harmful content.

Row type	Prompt	Observed behavior
Benign	Sensitive prompt preserved	Refusal boundary remains model-specific
Benign	Clean prompt preserved	Access is retained only at some operating points
Harmful	Prompt redacted	Original refusal preserved; enabling content redacted
Harmful	Prompt redacted	Original refusal preserved; enabling content redacted

F Base Model IDs and Release Plan

The base model IDs are `Qwen/Qwen2.5-3B-Instruct`, `Qwen/Qwen2.5-7B-Instruct`, `meta-llama/Meta-Llama-3-8B-Instruct`, `QCRI/Fanar-1-9B`, and `ALLaM-AI/ALLaM-7B-Instruct-preview`. The ALLaM Hugging Face page redirects to `humain-ai/ALLaM-7B-Instruct-preview`. The planned release includes training and evaluation code, configurations, seeds, judge prompts, anonymized audit labels, transformation or reconstruction scripts, and permitted adapter or checkpoint identifiers. No public artifact URL is available in the current record, so the paper does not claim that these materials are already released.