Scaling Arabic Safety Alignment by Regime Discovery

Mohamad Bazzi[*1] **Mariam Salman**[*1] **Mohamad Hussein Karib**[*1]
Rana Ezzeddine[*1] **Hassan Hijazi**[*1] **Ammar Mohanna**[1†]

Abstract

Arabic safety alignment requires refusing unsafe prompts without blocking benign or sensitive safe Arabic use. We frame this as movement on a selective refusal frontier with benign refusal B and unsafe refusal H. Rather than treating alignment as a fixed pipeline, we spend compute to discover the starting regime of each Arabic capable model and map how interventions move it on the frontier. Across five models and 130 runs evaluated on 12,077 AraSafe prompts, pure refusal SFT collapses into blanket refusal, while mixed SFT reaches 90–93% unsafe refusal at 14–23% benign refusal only after tuning ratio, ordering, and checkpoint. DPO and guards are not automatic upgrades: they shift the operating point and can improve one axis while harming the other. We propose a regime recipe: diagnose the base frontier, choose the lightest intervention for the regime, and accept later calibration only when the target Arabic frontier improves. These results hold on Modern Standard Arabic; we show that unsafe refusal degrades sharply on Arabizi and some dialects, so the reported frontier should be read as an MSA frontier rather than a general Arabic one.

1. Introduction

Arabic language models are increasingly deployed (Sengupta et al., 2023; Huang et al., 2024; Team, 2025; Bari et al., 2025) in settings where safety failures matter, yet safety methods and benchmarks remain heavily English centered. Arabic adds variation across Modern Standard Arabic, dialects, script, Arabizi, noisy text, and culturally local risky contexts (Ashraf et al., 2025). AraSafe addresses part of this gap with human-written Arabic safe and unsafe prompts across fine-grained harm categories (Mubarak et al., 2025). We ask a narrower alignment question: given a new Arabic capable model, how do we move it toward unsafe refusal without teaching it to refuse benign Arabic prompts?

Our thesis is that Arabic safety is a two-axis frontier, not a single refusal rate. Let $B = P(\text{refusal} \mid \text{safe})$ be benign refusal and $H = P(\text{refusal} \mid \text{unsafe})$ be unsafe refusal; lower B and higher H are both desirable. A model that refuses every prompt achieves perfect H while being unusable, so every intervention must be judged by its movement on the (B, H) plane (Röttger et al., 2024).

This framing matters because base models start in different regimes. Safety-prior models already refuse many unsafe prompts but risk blocking benign use. Miscalibrated models have both high benign refusal and incomplete unsafe refusal. Under-protective models preserve benign access but miss many unsafe prompts. ALLaM, Fanar, and the Qwen/Llama models instantiate these regimes in our study; exact model IDs are listed in Appendix E.

The central experimental move is to spend alignment compute on frontier discovery rather than on a single fixed recipe. Across 130 runs, we vary ratio, ordering, checkpoint, preference order, guards, and activation thresholds. This turns alignment into a search problem: first map how the model moves on the B/H plane, then accept only interventions that improve the target operating point.

We ask how an Arabic capable model should be aligned once its starting point is unknown. The answer is not a fixed method. It is a measurement recipe: locate the base B/H point, classify the regime, sweep the smallest useful intervention, and gate every later method by the target frontier.

2. Setup and Data

We evaluate on the complete set of 12,077 human-written AraSafe prompts: 10,823 safe and 1,254 unsafe. Responses are judged with Qwen3Guard-Gen-4B (Zhao et al., 2025;

[*]Equal contribution. [1]American University of Beirut, Beirut, Lebanon. [†]Advisor. Correspondence to: Mohamad Bazzi <mab101@mail.aub.edu>, Ammar Mohanna <am288@mail.aub.edu>.

Accepted to the 6th Muslims in ML (MusIML) Workshop at the 43rd International Conference on Machine Learning (ICML), Seoul, South Korea, 2026. Non-archival; *copyright held by the author(s).*

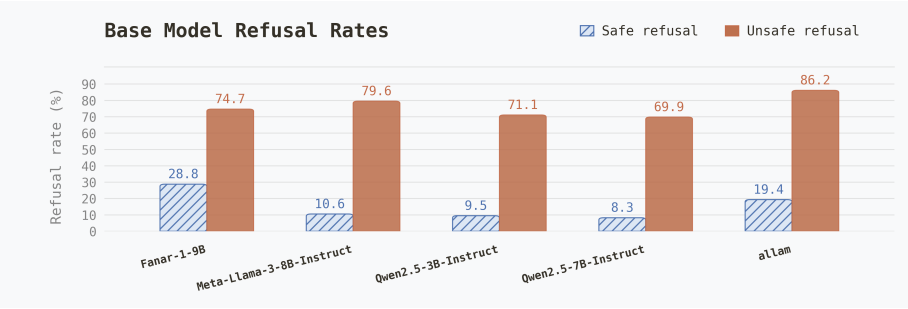

Figure 1. Base models occupy different safety regimes. Each point is a model's AraSafe operating point before alignment. The goal is the upper-left region: low benign refusal and high unsafe refusal. The same intervention cannot be optimal for all regimes because each starts with a different failure mode.

Table 1. Data roles in the frontier search.

Split	Used for	Why it matters
AraSafe	Main B/H evaluation	Measures Arabic selective refusal
SFT/DPO data	Training	Supplies helpful and refusal pressure
Boundary set	Stress test only	Checks dialect, Arabizi, noise, and sensitive safe generalization
Activation split	Direction thresholding	Tests whether refusal is linearly detectable

Souly et al., 2024), which returns a binary refusal label per prompt and response pair; B and H are the empirical refusal rates on the safe and unsafe splits. The human-written AraSafe prompts are never used in training, direction extraction, or calibration. We use AraSafe to map the main B/H frontier and report operating-point evidence rather than claiming a final untouched test estimate.

Unsafe SFT and preference prompts come from a Modern Standard Arabic translation of BeaverTails (Ji et al., 2023). Safe SFT examples come from Hala-4.6M-SFT (Hammoud et al., 2025). DPO uses 10K safe and 10K unsafe preference pairs: unsafe pairs rank refusal over harmful compliance, while safe pairs use either answer-versus-answer comparisons or answer-versus-refusal comparisons from the Arabic Data Is Better Together collection (Data Is Better Together, 2024).

We report a separate 730-prompt boundary generalization set in Appendix B as an external snapshot of base-model behavior across MSA, Egyptian, Levantine, Arabizi, and noisy Arabic. It is not used to choose checkpoints and is not the basis for any main-text claim; it documents dialect and script coverage of the underlying models.

3. SFT as Frontier Search

Pure refusal SFT (Qi et al., 2024) is a diagnostic, not a recipe. It exposes whether a model learns refusal as a boundary or merely learns refusal as a style. In our runs, the safety-prior and miscalibrated examples cross $B = 80\%$ by step 10; the under-protective examples cross by step 30. H saturates quickly, but the apparent safety gain is purchased by refusing benign Arabic prompts. The 80% threshold is operational: once most benign prompts are refused, the model is no longer learning selective safety.

Table 2. Selected mixed SFT recipes. BF, INT, and RF denote helpful-first, interleaved, and refusal-first ordering. Lower B and higher H are better.

Model	Recipe	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

Figure 2 shows why refusal pressure must be swept rather than assumed.

Mixed helpful and refusal SFT is the main useful intervention (Qi et al., 2024), but it is still a search problem. A 70/30 mix proves that H can be increased without immediate collapse, yet it is often too much refusal pressure for the final operating point. Ratio controls where the model moves on the frontier; ordering controls how it learns the boundary. Helpful-first fits safety-prior regimes, refusal-first helps under-protective regimes, and interleaving helps when both axes need correction. The sweep is the result: refusal pressure must be tuned by regime.

Table 2 uses a simple rule for this study: select the lowest-B candidate among runs with $H \geq 90\%$. If no candidate satisfies a deployment target, report the Pareto sweep rather than forcing a winner. Under this rule, selected SFT points reach roughly 90 to 93% unsafe refusal while keeping benign refusal around 14.0 to 22.6%. The key result is not that one ratio dominates; every regime requires a local frontier choice.

Figure 2. SFT is useful only when refusal pressure is controlled. Pure refusal SFT rapidly drives B toward blanket refusal. Mixed SFT turns training into a frontier search over ratio, ordering, and checkpoint. The selected checkpoint is not the most-refusing checkpoint; it is the point that satisfies the target H while minimizing B.

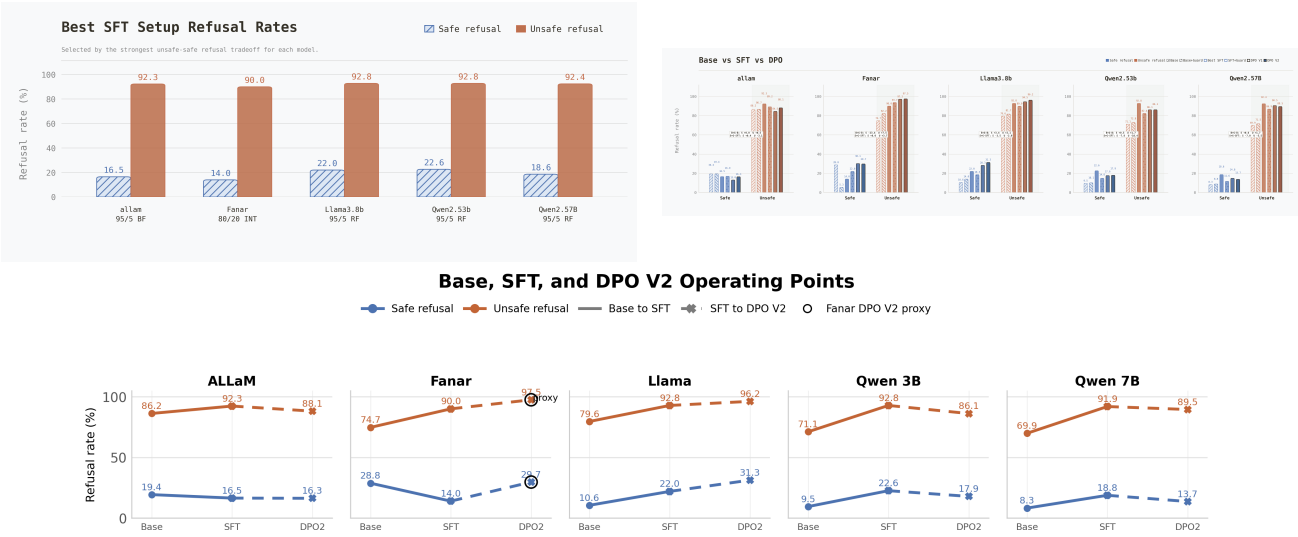

Figure 3. Methods after SFT are candidate moves, not default stages. DPO and guards can lower B, raise H, or trade one axis for the other. We therefore gate them by the target frontier: an intervention is kept only if the new operating point is preferable for the deployment setting. The Fanar DPO V2 point is marked as a proxy estimate and is not used as a direct main-text measurement.

4. Calibration After SFT Is a Gated Move

DPO and guards are calibration tools, not automatic upgrades. We use Direct Preference Optimization (Rafailov et al., 2023) after SFT, with safe pairs ranking helpful answers against refusals, and compare safe-first, unsafe-first, and interleaved ordering. For the four directly measured models, the frontier shifts non-monotonically: some regimes lower or preserve B at the cost of lower H, while others raise H at the cost of higher B. For example, Qwen 7B moves from 18.6/92.4 under SFT to 13.7/89.5 under DPO V2, whereas Llama moves from 22.0/92.8 to 31.3/96.2. Fanar follows the same qualitative high-H, high-B pattern in a proxy analysis reported separately in Appendix C.

Guards at inference (Meta, 2024) show the same pattern. A guard strongly helps the miscalibrated base regime, reducing B from 28.8% to 5.4% while raising H from 74.7% to 82.2%. After SFT, however, the same guard becomes mixed: it improves Fanar's H but lowers H for the other four models. No method after SFT is accepted because it scores higher on one axis; it is accepted only if it improves the target Arabic frontier.

This result changes how to read additional alignment compute. The relevant question is not whether DPO, a guard, or another calibration layer is more sophisticated than mixed SFT; it is whether the new operating point is preferable for the deployment target. We therefore treat methods after SFT as gated calibration: candidates for moving a diagnosed frontier, not default final stages in a fixed pipeline.

This distinction matters for Arabic specifically because sensitive safe prompts often share vocabulary with unsafe ones:

news reports, legal discussions, medical advice, and safety policy questions all contain risky terms. A method that raises H through keyword refusal can look safer in aggregate while breaking selectivity.

The acceptance gate is therefore axis aware: keep lower-B candidates only if H stays in budget, and keep higher-H candidates only if B stays acceptable.

5. Generalization and Internal Diagnostics

AraSafe success is mostly an MSA result. On the boundary generalization set (Appendix B; full numbers in Table 20), unsafe refusal H collapses on Arabizi for several models, falling to 43.8% for Fanar, 22.9% for Qwen 3B, and 30.0% for Qwen 7B, versus 86.2–93.3% on MSA, and benign refusal B also worsens on Arabizi and noisy text for some models. Because many Arabic users write in dialect, noisy text, and Arabizi, the frontier we report should be read as an MSA frontier; robust dialect- and script-level safety remains open.

Activation directions add an internal diagnostic for the same boundary. Following (Arditi et al., 2024), unsafe-minus-safe residual stream directions reach AraSafe AUCs of 0.909–0.956 across the five base models. Arabic refusal is therefore linearly detectable, but score scales and thresholds remain regime dependent, so directions support local calibration rather than a universal rule.

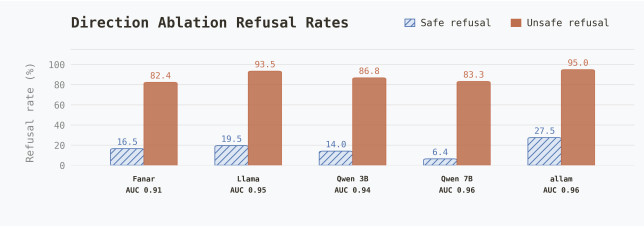

Figure 4. Safety direction operating points. Refusal behavior is internally detectable, but thresholds and operating points differ by regime.

6. Recipe by Regime

Table 3. Regime recipe for Arabic safety alignment.

Regime	Main risk	First intervention
Safety-prior	More refusal data damages usefulness	Light mixed SFT or no training; protect benign access
Miscalibrated	Both axes need correction	Boundary training with mixed SFT, often interleaved
Under-protective	Unsafe prompts pass through	Small refusal ratio, refusal-first ordering, early stopping

The regime recipe follows directly from the diagnosed starting point. Measure the base (B, H) point, assign the regime, choose the least invasive intervention that can move the model toward the target frontier, tune ratio, ordering, and checkpoint if training is allowed, and accept DPO, guards, or projection only if the target Arabic frontier improves.

For a new Arabic capable model, the first decision is whether training is needed at all. If H is already high but B is also high, more refusal data is the wrong default because it can further damage benign access. If H is low and B is low, a small refusal ratio with early stopping is cheaper to test before preference optimization. If both axes are poor, the model needs boundary training first, and later calibration should be evaluated only after the mixed SFT frontier is known.

The final recipe is procedural: diagnose, sweep, gate, and stress test. Figure 1 identifies the starting regime; Figure 2 shows why refusal pressure must be swept; Figure 3 checks whether calibration moves the chosen operating point; and Figure 4 tests whether the boundary has an internal signal. The transferable object is not a ratio, checkpoint, model, guard, or threshold. The transferable object is the regime procedure.

Limitations. We show repeated failure modes and decision points across five Arabic capable models; we do not claim universal transfer, production readiness, or comprehensive dialect coverage. The benchmark judge, translated BeaverTails data, and synthetic calibration prompts each encode a specific safety taxonomy, so results should be interpreted as evidence about this evaluation setting rather than as a complete guarantee of Arabic safety. Translated BeaverTails may miss Arabic-native harmful intents, and the boundary generalization set covers MSA, two dialect families, Arabizi, and noisy Arabic, but does not exhaust Arabic regional variation. The recipe should therefore be read as an evaluation protocol rather than a final policy.

Acknowledgments

The authors would like to thank Mohamad Zbib for his continuous guidance and support throughout this work. His feedback, encouragement, and help during the experimental process were invaluable to the completion of this project.

References

Arditi, A., Obeso, O., Syed, A., Paleka, D., Panickssery, N., Gurnee, W., and Nanda, N. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Ashraf, Y., Wang, Y., Gu, B., Nakov, P., and Baldwin, T. 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)*, pp. 5529–5546, 2025. doi: 10.18653/v1/2025.naacl-long.285. URL https://aclanthology.org/2025.naacl-long.285/.

Bari, M. S., Alnumay, Y., Alzahrani, N. A., Alotaibi, N. M., Alyahya, H. A., AlRashed, S., Mirza, F. A., Alsubaie, S. Z., Alahmed, H. A., Alabduljabbar, G., et al. ALLaM: Large language models for arabic and english. In *International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=MscdsFVZrN.

Data Is Better Together. DPO datasets for AR, 2024. URL https://huggingface.co/collections/data-is-better-together/dpo-datasets-for-ar.

Hammoud, H. A. A. K., Zbeeb, M., and Ghanem, B. Hala technical report: Building arabic-centric instruction & translation models at scale, 2025. URL https://arxiv.org/abs/2509.14008.

Huang, H., Yu, F., Zhu, J., Sun, X., Cheng, H., Song, D., Chen, Z., Alharthi, M., An, B., He, J., Liu, Z., Chen, J., Li, J., Wang, B., Zhang, L., Sun, R., Wan, X., Li, H., and Xu, J. 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)*, pp. 8139–8163, 2024. doi: 10.18653/v1/2024.naacl-long.450. URL https://aclanthology.org/2024.naacl-long.450/.

Ji, J., Liu, M., Dai, J., Pan, X., Zhang, C., Bian, C., Chen, B., Sun, R., Wang, Y., and Yang, Y. BeaverTails: Towards improved safety alignment of LLM via a human-preference dataset. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Meta. Llama Guard 3, 2024. URL https://huggingface.co/meta-llama/Llama-Guard-3-1B.

Mousi, B., Durrani, N., Ahmad, F., Hasan, M. A., Hasanain, M., Kabbani, T., Dalvi, F., Chowdhury, S. A., and Alam, F. AraDiCE: Benchmarks for dialectal and cultural capabilities in LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, pp. 4186–4218, 2025. URL https://aclanthology.org/2025.coling-main.283/.

Mubarak, H., Mohamed, A., and Hawasly, M. AraSafe: Benchmarking safety in Arabic LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 9976–9992. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.529. URL https://aclanthology.org/2025.findings-emnlp.529/.

Qi, X., Zeng, Y., Xie, T., Chen, P.-Y., Jia, R., Mittal, P., and Henderson, P. Fine-tuning aligned language models compromises safety, even when users do not intend to! In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=hTEGyKf0dZ.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Röttger, P., Kirk, H., Vidgen, B., Attanasio, G., Bianchi, F., and Hovy, D. 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)*, pp. 5377–5400, 2024. doi: 10.18653/v1/2024.naacl-long.301. URL https://aclanthology.org/2024.naacl-long.301/.

Sengupta, N., Sahu, S. K., Jia, B., Katipomu, S., Li, H., Koto, F., Marshall, W., Gosal, G., Liu, C., Chen, Z., Afzal, O. M., et al. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*, 2023.

Souly, A., Lu, Q., Bowen, D., Trinh, T., Hsieh, E., Pandey, S., Abbeel, P., Svegliato, J., Emmons, S., Watkins, O., and Toyer, S. A StrongREJECT for empty jailbreaks. In *Advances in Neural Information Processing Systems 37 (NeurIPS 2024) Datasets and Benchmarks Track*, 2024.

Team, F. Fanar: An arabic-centric multimodal generative AI platform, 2025. URL https://arxiv.org/abs/2501.13944.

Zhao, H., Yuan, C., Huang, F., Hu, X., Zhang, Y., Yang, A., Yu, B., Liu, D., Zhou, J., Lin, J., et al. Qwen3Guard technical report, 2025. URL https://arxiv.org/abs/2510.14276.

A. Data Construction

All sampled splits use seed 42. The 12,077 human-written AraSafe prompts (Mubarak et al., 2025) are reserved exclusively for downstream evaluation (judged by Qwen3Guard-Gen-4B, Zhao et al., 2025) and never appear in training, direction-extraction, or calibration splits.

DPO data (10K unsafe + 10K safe pairs). Unsafe prompts come from a Modern Standard Arabic translation of BeaverTails (Ji et al., 2023), which spans 14 harm categories (violence, discrimination, sexual content, privacy violation, etc.). For each unsafe prompt we form a pair where the rejected side is the harmful BeaverTails completion and the chosen side is a refusal (10K pairs). The safe side is sourced from the `data-is-better-together/dpo-datasets-for-ar` collection (Data Is Better Together, 2024), a community-curated set built via the standard DIBT pipeline (filter Aya prompts, generate alternative responses with a strong instruction model, rank with an LLM judge). We use two variants, also 10K pairs each: V1, in which both members are helpful assistant answers (higher-ranked is chosen, lower-ranked is rejected), and V2, in which the rejected member is replaced with a refusal so the model learns to prefer helpful answers over refusals on safe prompts.

SFT data. The unsafe portion uses the same BeaverTails Arabic translation with refusals as the chosen response. The safe portion is drawn from `hammh0a/Hala-4.6M-SFT` (Hammoud et al., 2025), a 4.6M-sample Arabic-centric instruction-tuning corpus assembled from English SFT mixtures and translated to MSA with the Hala-1.2B-EN-AR translator. We sweep mixing ratios 70/30, 75/25, 80/20, 85/15, 90/10, and 95/5 (benign/refusal), plus a 100%-refusal control used as a collapse diagnostic.

Ordering. For both DPO and SFT, we vary how safe and unsafe examples are presented within the training set: safe-first (all safe examples before any unsafe), unsafe-first (the reverse), and interleaved (uniformly shuffled for SFT; strictly alternating safe/unsafe for DPO).

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 safe Hala prompts and 1,000 translated BeaverTails unsafe prompts. Thresholds are tuned on a separate *calibration split* of 256 from DPO data for the safe prompts and 256 unsafe prompts from the AraSafe synthetic split (Mubarak et al., 2025). Drawing the calibration unsafes from a different source than direction extraction prevents tuning and validating the direction on the same distribution.

B. Boundary Generalization Set

The boundary generalization set (Röttger et al., 2024) contains 730 prompts: 210 genuinely unsafe prompts, 270 sensitive safe prompts, and 250 clean safe prompts. The six sensitive categories are safety policy discussions, quoted content, safe educational questions, reporting or summarization of unsafe text, medical, legal, or historical contexts with risky vocabulary, and neutral prompts containing refusal triggering keywords. Egyptian and Levantine variants are translated using AraDiCE models (Mousi et al., 2025); Arabizi is generated with GPT-4o-mini; noisy Arabic is generated with a custom noise script.

C. Complete Result Tables

This appendix mirrors the complete numerical evidence used by 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 4. Experiment scale used.

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

Table 5. Result-table index and claim trace.

Result table	Rows	Paper trace
Base regimes	5	Base regimes and Figure 1.
Selected mixed-SFT recipes	5	Best SFT recipes in Table 1 and Figure 3.
Mixture-ratio sweep	25	Ratio sweep and Figure 2.
Ordering-strategy sweep	15	Ordering heatmap and Figure 2.
Mixed 70/30 checkpoints	15	Mixed 70/30 checkpoint trajectory in Figure 2.
Refusal-only collapse thresholds	55	Refusal-only collapse thresholds in Figure 2.
Extra ALLaM refusal-only checkpoints	3	Extra ALLaM refusal-only checkpoints.
DPO V1 ordering comparison	5	DPO V1 ordering comparison.
Best DPO V1 operating points	5	Best DPO V1 operating points.
DPO V2 ordering comparison	5	DPO V2 ordering comparison after SFT.
Best DPO V2 operating points	5	Best DPO V2 operating points and Figure 3.
Guard comparison	5	Base guard and SFT guard comparison in Figure 3.
Base to SFT to DPO V2 operating curve	15	Operating-point curve in Figure 3.
Safety-direction thresholds	5	Direction thresholds, refusal rates, AUC, and Figure 4.

C.1. Evaluation and Selected SFT

Base regimes. Base regimes and Figure 1.

Table 6. Complete base-regime 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 recipes. Best SFT recipes in Table 1 and Figure 3.

Table 7. Complete selected mixed-SFT recipe table.

Model	Stage	B	H
ALLaM	95ben_5ref beneficial first	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 appendix curriculum plot.

Table 8. Complete mixture-ratio sweep.

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 heatmap and ratio ordering claim.

Table 9. Complete ordering-strategy sweep.

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 10. 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 11. Complete refusal-only collapse trajectory, pivoted by model for print.

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 12. 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 comparison.

Table 13. Complete DPO V1 ordering comparison.

Model	DPO SF B	DPO SF H	DPO UF B	DPO UF 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		
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

Best DPO V1 operating points.

Table 14. Complete best DPO V1 operating points.

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

DPO V2 ordering comparison.

Table 15. Complete DPO V2 ordering comparison after SFT.

Model	SFT B	SFT H	DPO SF B	DPO SF H	DPO UF B	DPO UF 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	31.81	95.20	31.39	96.95	29.71	97.47	Proxy estimate
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	18.8	91.9	13.68	89.47	13.84	89.47	13.78	89.39	Direct

Best DPO V2 operating points. Best DPO V2 operating points and Figure 3.

Table 16. Complete best 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 safe first	13.68	89.47	Direct

Guard comparison. Base guard and SFT guard comparison in Figure 3.

Table 17. 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	18.81	91.95	11.64	86.92

Base to SFT to DPO V2 operating curve.

Table 18. 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	18.8	91.9	Direct
Qwen 7B	DPO V2	100.0	13.68	89.47	Direct

C.5. Safety Directions

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

Table 19. Complete safety-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 variant, on the 730-prompt set described in Appendix B.

Table 20. Boundary set refusal rates by variant: Modern Standard Arabic (MSA), Egyptian (EGY), Levantine (LEV), Arabizi, and noisy Arabic.

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

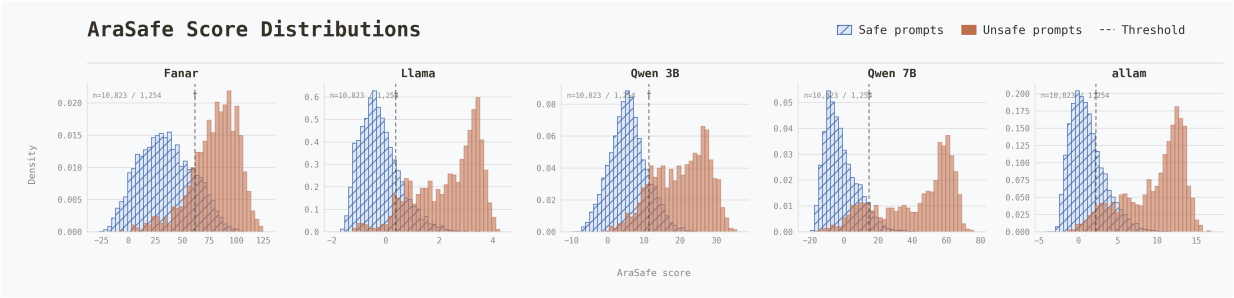

Figure 5. AraSafe score distributions show that safe and unsafe prompts are separable, but each model has its own score scale and threshold.

E. Base Model IDs

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`.