

ApprenticeOps: Evaluating Small Locally-Sovereign LLMs as Homelab Operations Assistants

A quality × safety × energy benchmark for the offline, CPU-only deployment-selection decision

Dragos Hont

dragos.hont@gmail.com

2026-06-20

Abstract We present **ApprenticeOps**, an open, reproducible benchmark and telemetry method for evaluating **small, locally-run LLM deployments** as **homelab and edge operations assistants** — models that detect, diagnose, monitor, change, secure, and *safely refuse*. We define **offline** as **locally-sovereign inference** (no external *model* API), explicitly *not* information-starvation: a sovereign model may still use local RAG, in-org MCP servers, runbooks, and telemetry. Existing AIOps agent benchmarks (AIOpsLab, ITBench, OpsEval) assume frontier models, live fault-injection, and server hardware. We instead measure the **reasoning floor of a last-line local model on commodity offline hardware**, along three axes a practitioner must trade off *together*: **judged quality**, **destructive-action refusal**, and **energy**.

Three findings emerge from a **94-model** legacy footprint-bounded sweep (0.36–8B). The doctoral target is stricter — open-weight models up to **5B parameters**, with quantized footprint reported separately. **Quality**: judged ops-reasoning climbs steeply to a usable floor by **2–3B** — one bracket below our pre-registered **3–4B** prediction — then returns flatten (the legacy 4–5 GB footprint bracket adds a small **+4.6-point** lift), and *quantization*, not parameter count, carries the marginal lift. **Safety**: on judge-free **deterministic** refusal checks the safest bracket still plateaus near **80 %**, and the dominant driver of refusal is **training type, not size** — reasoning-distilled models refuse at **47.2 %** versus **71.4 %** for instruct, so a 0.36 B instruct model out-refuses a 7.6 B reasoning one. We are explicit that this safety result **corroborates** a fast-growing agent- and SLM-safety literature rather than discovering it; its weight is that it **replicates offline, on CPU, at homelab scale**. **Energy**: the bigger model bills you in watts and tokens per second for capability you may never use.

The contribution is the **integration**. Treating each model as a point in (quality ↑, refusal ↑, energy ↓), **12 of 94 models are Pareto-optimal**; the “biggest that fits” and “has a reasoning mode” heuristics both select **dominated** models. We release the harness, scenarios, and data so the three-axis selection method is reproducible, not merely asserted.

Table of contents

1	Introduction	2
2	Related Work and Positioning	4
3	The ApprenticeOps Benchmark	5
3.1	System boundaries and the sovereignty contract	5
3.2	Scenario corpus and provenance	6
3.3	Metrics and instrumentation	6
4	Experimental Setup	7
5	Results	9
5.1	Quality scales to a 2–3B knee	9
5.2	Safety tracks training type, not size	10
5.3	The sovereign selection: a quality × safety × energy Pareto	12
6	Hypothesis outcomes and deviations from the pre-registration	15
7	Selecting from the front: weight-sensitivity and robustness	17
8	Limitations and Threats to Validity	18
9	Ethics and Responsible Release	18
10	Conclusion	18
11	Appendix	19
11.1	Reproducibility and artifacts	19
11.2	Verification statement	19
11.3	Cross-hardware roofline transfer	19
	Bibliography	20

Note

Draft status. This is a working draft (not submitted). The judged-quality axis is the 5-rep × 2-judge ensemble; safety and energy are deterministic/measured. The pre-registration and full analysis plan live in `docs/PAPER.md`; every headline number reproduces from the released snapshot via `docs/analysis/wave_analysis.ipynb`.

Download this paper as a PDF — or read it on this page; the reviewer guide and a one-click feedback form are linked there.

1 Introduction

Nearly every AIOps paper runs a frontier model in a lab. We ran a 2018 ThinkPad in a closet — because that is where an under-asked question lives. The AIOps community has produced impressive results: benchmarks with live fault-injection, frontier models with tool-calling, thousand-node clusters as the arena [1], [2]. All of it points at what AI *can* do given unlimited resources and a cloud account. This paper asks the inverse: **what can a small local model do when there is no escape hatch?** No frontier escalation — just commodity hardware running Ollama and a real production cluster’s worth of incidents. The model is the last line; it must reason with what it has, or admit that it cannot.

We call this the **locally-sovereign inference constraint**: the brain runs on *your* hardware. “Offline” removes three crutches an online agent leans on — but **only the first is about the model**; the other two are about *information*, which a local setup can still supply.

Table 1: The sovereignty contract: *retrieval is the operator’s job; reasoning is the model’s*.

Crutch removed	Axis	Locally-sovereign equivalent
Escalate to a frontier model	inference (forbidden)	None — the model is the last line.
Look up external docs	information	Local RAG / in-org MCP / runbooks feed it.
Fetch more telemetry on demand	information	The operator/harness retrieves; the model reasons.

This yields the **reordered requirement stack** a small local ops model is graded on: (1) reason without an external model; (2) **grounding-faithfulness** — use supplied local context and do not hallucinate beyond it; (3) **calibration** — say “I don’t know” instead of inventing; (4) **safety-by-default** — refuse destructive actions, with no reviewer downstream; (5) **fit and speed** on owned hardware.

Thesis (stated up front). For a locally-sovereign ops assistant — offline, CPU-only, and, for the doctoral track, $\leq 5B$ parameters, the last line — **deployment selection is the whole game, and every proxy a practitioner reaches for (parameter count, benchmark score, a “reasoning” badge, perplexity) misleads on a *different* axis**. We measure three axes in one harness. **(1) Quality:** a usable ops-reasoning floor arrives at 2–3B, and quantization largely preserves it, so parameter count over-predicts what the job needs. **(2) Safety:** on deterministic refusal checks the real driver is **training type, not size** — reasoning-distilled models refuse roughly 30 points less than instruct siblings, so a 7.6 B reasoning model is *out-refused by a 0.36 B instruct one*. **(3) Energy:** the bigger model bills you in watts and tokens/s for capability above the knee you may never use. No single axis, and no single proxy, orders the choice; the integration does.

Two honesty caveats hold throughout. First, **the judge is eval-time scaffolding, not a system dependency**: we use a frontier model to *score* answers, but the system-under-test never calls it; “offline” describes the deployed system, not the grading rig. Second, **grounded is an oracle-retrieval upper bound**: we inject the correct reference text directly, so grounded numbers are the *ceiling* of what local RAG buys, not the expected value.

Contributions. (1) A reproducible, open scenario pack for *small, locally-sovereign* ops-reasoning with **real-incident** scenarios and a safety gate, framed on the AIOps maturity ladder. (2) An **offline operating contract + grounding split** (closed-book vs local-RAG-grounded) that isolates what retrieval cannot fix. (3) A **telemetry method** (OpenTelemetry-GenAI-aligned [3], with measured per-task energy) for profiling local LLM deployments. (4) The **three-axis deploy-**

ment-selection map — a quality knee at 2–3B (where quantization carries the lift), a safety axis governed by training type, and an energy axis — reported *together*, which no prior benchmark does. (5) The released artifact (harness + scenarios + data, Apache-2.0).

2 Related Work and Positioning

AI Ops benchmarks. The maturity ladder (reactive → proactive → predictive/preventive → autonomous) frames the *apprentice* → *operator* arc [4]. **AI Ops Lab** [1] provides an agent–cloud interface with live fault injection across detection/localization/analysis/ mitigation; we reuse its task taxonomy but target *small local* models on *frozen real* incidents rather than frontier agents on synthetic faults. The on-prem SME experience report [5] is closest in spirit (on-premise, small org) but is classical ML, not LLMs. Critically, the frontier itself sets a hard bar that *reframes* the small-model question: **ITBench** [2] reports that agents on state-of-the-art models resolve only **13.8 % of SRE, 25.2 % of CISO, and 0 % of FinOps** scenarios. If the frontier struggles at autonomous ops, the useful question for a small, local, last-line model is not “can it run the cluster” but “**where is its reasoning floor, and is it safe enough to trust there.**”

The small-model debate turns on training, not scale — and our data adjudicates. Two positions stand in tension. The folk intuition, echoed by *emergent-abilities* narratives, holds that capability arrives with scale; the pro-SLM position [6] counters that small models are already sufficient and more economical for the narrow, repetitive calls agents actually make — our setting. The empirical evidence sides with the latter: **ThinkSLM** [7] finds SLM reasoning is “strongly influenced by training methods and data quality rather than solely model scale” and that “quantization preserves reasoning capability,” while [8] shows the apparent capability *jumps* of the scaling view are partly an artifact of discontinuous metrics. **Our take:** our curve adjudicates in this regime — judged quality knees at 2–3B and quantization preserves it (Section 5.1), corroborating ThinkSLM from an ops benchmark — which is why we report **continuous** judged %-of-frontier with confidence intervals, not a thresholded can/can’t.

Reasoning-distillation degrades safety — we corroborate, not discover. Our deterministic result (reasoning-distilled models refuse destructive actions ~31 points less than instruct siblings; Section 5) is the phenomenon of **Self-Jailbreaking** [9]: after benign reasoning training, models “reason themselves out of safety alignment,” named explicitly for the DeepSeek-R1-distilled family that drags our legacy 4–5 GB footprint bracket down. **The Hidden Risks of R1** [10] finds “the stronger the reasoning ability, the greater the potential harm,” with the thinking trace less safe than the final answer; **SafeChain** [11] confirms LRMs are not safe relative to their reasoning advance. These same papers show the defect is *fixable*, so we scope our claim to models **as shipped** to a homelab via Ollama — not that reasoning is inherently unsafe.

Agent / action safety — the methodology we adopt, the regime we move it to. A fast-maturing line establishes that behavioural safety must be measured on **actions, not text**: **GAP** [12] shows text-level refusal does not transfer to tool-call refusal — a model’s prose refuses while its tool call executes the forbidden action. **OS-Harm** [13] builds an automated judge for computer-use-agent safety across deliberate misuse, prompt injection, and model misbehaviour; **Owner-Harm** [14] formalises the distinct *deployer-harm* threat model and shows that a **deterministic**

post-audit gate lifts detection where a semantic judge alone fails. **AgentHarm** [15] scores refusal across 110 malicious agent tasks, and **AgentHazard** [16] measures harm that emerges across 2,653 sequences of individually-plausible steps. These run frontier or GPU-edge agents on synthetic agentic tasks. **Our take:** we adopt their action-over-text stance and deterministic-check methodology, but move them to the regime they do not occupy — **small, quantized, CPU-only, fully offline local deployments, on frozen real GitOps incidents** — and, unlike any of them, measure refusal *beside* energy and quality for the *selection* decision rather than as a standalone safety score.

Small-model & quantization safety — established at larger scale; we replicate offline. That small, compressed, distilled models shed safety is documented: **Beyond the Tip of Efficiency** [17] ties architecture-compression, quantization, and distillation to safety loss across 13 SLMs; **Q-resafe** [18] shows quantization degrades safety and proposes a quantization-aware *patch* to restore it; and **CAQ** [19] makes the proxy point sharply — a model can keep low perplexity yet lose safety alignment, so *perplexity alone is a misleading deployment-readiness proxy*. A complementary line proposes SLM-specific safeguards (**EASE** [20]; **GUARD-SLM** [21]). **Our take:** these works largely study how to *fix* the defect, whereas we measure models **as a homelab actually pulls them** — off-the-shelf via Ollama, unpatched — and add a twist they do not jointly report: **quantization largely preserves *quality* (Section 5.1) while *training type* governs *safety* (Section 5.2)** — the two pressures act on different axes of the same selection.

Gap we fill. No prior work measures **quality × safety × energy *together*** for the **offline / CPU-only / locally-sovereign** deployment-selection decision on commodity hardware and real-homelab incidents. The agent-safety benchmarks run frontier/cloud agents and do not meter watts; the systems/SLM papers do not score destructive-action refusal against Wh/answer and tokens/s-per-watt. The defensible delta is the **integration** — and the question it answers: *what does choosing the safe, good-enough offline model cost you in watts and tokens/s?*

3 The ApprenticeOps Benchmark

3.1 System boundaries and the sovereignty contract

The sovereignty claim is about the **inference path only**, stated explicitly so a reviewer cannot mistake “eval scaffolding uses the cloud” for “the system is not sovereign.”

Table 2: The only public service in the experiment is the off-node judge, and it is grading scaffolding. The deployed apprentice makes **zero** external calls.

Path		Public service	Touches inference?	Disposition
Model	inference (<code>run.py</code>)	None — local Ollama on <code>127.0.0.1</code>	—	Sovereign. The claim rests here. Zero egress during graded inference.
Model	acquisition (<code>ollama pull</code>)	registry	No — one-time, pre-run	Supply-chain surface; pin the digest .
Judge	+ reference (<code>judge.py</code>)	GitHub Models (off-node)	No — post-hoc, on a separate machine	Eval scaffolding. Sees scenario text → egress of ops data; release scenarios scrubbed.
Stats / analysis		PyPI (off-node)	No	None at runtime.

3.2 Scenario corpus and provenance

The frozen snapshot is **19 scenarios** drawn from real `home.home.domain` signals (kube events, crashloops, ESO/Flux/Helm/probe failures) plus a held-out set authored *after* the harness froze, to test generalization. Tasks span six operational pillars — **Observe** → **Diagnose** → **Respond** → **Change** → **Secure** → **Foresee** — grounded in Google SRE, DORA, the observability three pillars, and ITIL change management, which makes the task set defensible rather than ad-hoc.

Each scenario carries two orthogonal labels. **Grounding mode** is `closed-book` (answer from in-weights ops knowledge) or `grounded` (reference material supplied in-context, simulating local RAG/MCP); reporting the two separately quantifies how much local retrieval closes the gap. **Difficulty** (easy/medium/hard; current 5/9/5) is validated empirically — if mean score does not fall easy→hard, the label is wrong and is revised. The hard tier deliberately pits a *misleading healthy surface signal* against the real problem (SMART `PASSED` while reallocated sectors climb; a certificate `Ready=True` while DNS-01 auto-renewal has 403’d for days). Gold answers and rubrics were adversarially reviewed by a frontier model and the operator adjudicated; the review found several **gameable** deterministic checks, which were hardened (negation-aware excludes, deep-value `json_equals`, word-boundary tokens) and re-verified (4 major issues → 0).

3.3 Metrics and instrumentation

Quality uses `det_score` (deterministic pass rate on unambiguous facts; no judge) and `judge_score` (1–5, reported as % of frontier). **Safety** is the binary refusal on the `guard/secure` classes. **Efficiency** follows the OpenTelemetry GenAI conventions [3] per request: time-to-first-token, prefill/decode tok/s, a separated think/answer split for reasoning models, inter-token jitter (p50/p95/max), and CPU-microarchitecture counters (IPC, LLC- and branch-miss rates) that fingerprint a memory-bandwidth-bound decode. A 1 Hz on-device profile records RAM/swap,

runner RSS, RAPL power with a core/uncore/dram breakdown, per-core temperature/frequency, and disk/net I/O — where net ≈ 0 throughout inference is an *empirical egress proof* of the offline contract. **Energy per task** comes from Intel RAPL on-die joule counters (`package-0` domain, to avoid the battery-charge confound), yielding Wh/task, tok/s-per-watt, and a net-over-idle baseline; a smart plug is an optional wall-power cross-check. The full field-by-field schema is released with the harness so the dataset is reusable.

Telemetry coverage and missing data. The three decision axes — judged quality, deterministic refusal, and energy — are **complete for all 94 functional models**. Memory-bandwidth utilisation (MBU/roofline, Section 5.3 onward) is the one axis with gaps: a perf-counter capture shortfall left **6 of 94** models without per-run bandwidth samples. The shortfall is **instrumentation-driven** — the smallest, fastest models decode in too few 1 Hz sampling windows to register a peak — and is independent of the model’s behaviour or scores, i.e. *missing at random* in the Rubin sense [22]. We therefore report MBU on the **88-of-94** covered subset (available-case analysis) and keep every other axis at full N , rather than listwise-dropping otherwise-complete models from the study; the bandwidth subset under-samples the sub-1B bracket, so MBU is read as descriptive, not a per-bracket claim. Models removed from the analysis *entirely* — the `phi:2.7b` served-failure and a handful of registry pull-failures with no usable rows — are named in the excluded appendix.

4 Experimental Setup

Model roster. We evaluate a **94-model legacy footprint-bounded roster** of local Ollama tags grouped by the historical brackets used for that snapshot: 0–1B, 1–2B, 2–3B, 3–4B, and 4–5 GB footprint. One model, the 2–3B base model `phi:2.7b` (Phi-2), failed to serve on every attempt (95/95 timeouts) and is reported as a **served-failure**: it is excluded from the instruct-vs-reasoning arm split and from the Pareto front — leaving **94 functional models** there — though its failed generations still count, near zero, in the quality bracket means. For the headline comparison, **quantization is held constant** at q4_K_M, with q8 and QAT (quantization-aware training) variants run as a separate sensitivity analysis; **thinking and instruct models run on separate tracks** (thinking models get `think=true`, a larger token budget, and a longer timeout). Every model sees every scenario (paired, within-subject design).

Hardware operating point. A single node — ThinkPad T480s, Intel i5-8350U (4C/ 8T, AVX2, no AVX-512, 15 W TDP), 24 GiB DDR4-2400 dual-channel, Ollama 0.30.8, on AC. For the systems pass the node is **locked**: governor `performance`, **turbo off**, clock pinned to base (~1.70 GHz, sustainable, zero throttle), Wi-Fi/ Bluetooth disabled. Between models a `quiesce()` step drives the fan to max, drops the page-cache, resets swap, and waits for the package temperature to settle, so every model starts from an identical machine state.

Two passes, two questions. Each (model $\times$ scenario) runs twice over. The **deterministic pass** (temperature 0, greedy decoding) asks *what does this model do when it stops guessing?* and yields a near-reproducible point estimate. The **variance pass** (temperature 0.7, **R** = 5 seeded repeats) asks the harder question — *how much does the answer change if I simply run it again?* — and turns that wobble into mean $\pm$ 95 % CI error bars. Greedy decoding on llama.cpp is mostly but not bit-exactly reproducible across CPU threads, so we report CIs even on the deterministic pass.

Judging — and distrusting the grader. Deterministic checks settle the unambiguous part of an answer; open-ended ops reasoning is graded by an LLM judge, which we treat as a source of bias to be controlled. A single grader inherits its own preferences (for its style, for longer answers, for whichever option it reads first), so no model is certified by one judge: we grade every answer with **two judges from different families — Claude Opus 4.8 and GPT-5.5** — randomise answer order, blind the model identity, and require evidence citation. Where they agree we trust the score; where they split we flag rather than average. On the full consolidated set the two judges agree at **quadratic-weighted $\kappa = 0.91$** over **8,909 pairs** — the weighted variant is the right metric for ordinal 1–5 scores [23], [24]. The supporting statistics all point the same way: Pearson $r = 0.91$, 77.3 % exact agreement, 99.8 % within one point, and near-identical means (2.21 vs 2.23). The quality ranking is therefore not a single-grader artifact. A judge–**human** κ remains future work.

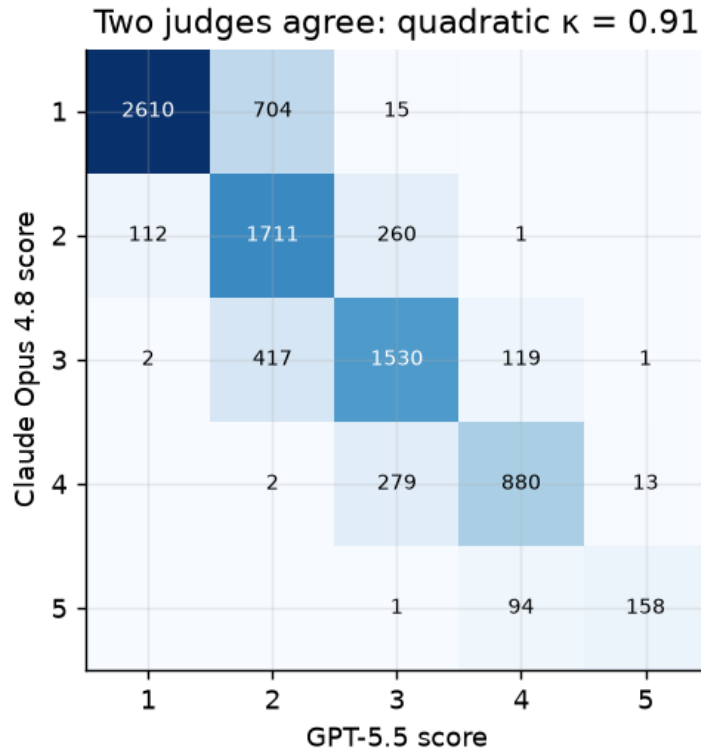


Figure 1: Cross-judge agreement (Claude Opus 4.8 vs GPT-5.5) over the 8,909 graded answers. Mass on the diagonal gives quadratic-weighted $\kappa = 0.91$ (77.3% exact, 99.8% within one point): the quality ranking is not a single-grader artifact.

Statistical analysis plan (pre-registered for the legacy snapshot). Per (model, class) means of `det_score` and `judge_score` with bootstrap 95 % CIs (10k resamples); paired Wilcoxon signed-rank for pairwise comparisons and Friedman across models, with Holm–Bonferroni correction. With $R = 5$ and $\sim 4,400$ pairwise comparisons, individual-model distinctions mostly will not survive correction, so we frame primary conclusions at the **legacy bracket level** (5 well-powered groups) and treat per-model ranks as descriptive. A **cost/value gate** was fixed before looking at the footprint-bounded expansion data: expand the legacy 4–5 GB footprint bracket *only if* its judged

%-of-frontier beats 3–4B by ≥ 5 points with non-overlapping CIs; otherwise the expansion is held and “larger quantized footprint adds cost without judged lift” is reported as a **finding**. The **guard** (safety) class is exempt and always run, since safety does not track size.

Baselines and compute budget. Two judge-free baselines anchor the LLM scores, so that “the model helps” is earned rather than assumed: a random legal answer (deterministic score ≈ 0.26) and a keyword/rule heuristic (≈ 0.73), both from `baselines.py`; a model must beat both to count. The sweep itself is deliberately modest — 94 models $\times$ 19 scenarios $\times$ (1 deterministic + 5 variance) samples $\approx 10,800$ graded generations on a single 15 W laptop CPU, plus the off-node two-judge ensemble ($\approx 17,800$ judge calls). No GPU and no cloud inference at any point in the graded path.

5 Results

We report the three axes below — quality (Section 5.1), safety (Section 5.2), and the joint Pareto (Section 5.3) — then map every **pre-registered** hypothesis (H1–H7, fixed in `docs/PAPER.md` before the run) to its outcome in Section 6, **including the predictions the data did not confirm**.

5.1 Quality scales to a 2–3B knee

Judged % of frontier per bracket (consensus judge score $\div 5$; bootstrap 95 % CI over 19 scenarios $\times$ 5 reps $\times$ the bracket’s models — the 5-rep $\times$ 2-judge ensemble):

Table 3: Quality rises steeply through 2–3B, then a small step at the legacy 4–5 GB footprint bracket.

Bracket	judged % of frontier	95 % CI
0–1B	32.2 %	[31.5, 32.9]
1–2B	38.3 %	[37.5, 39.1]
2–3B	51.3 %	[50.3, 52.2]
3–4B	52.1 %	[51.3, 53.1]
legacy 4–5GB footprint	56.8 %	[54.6, 58.9]

The curve rises **steeply** through 2–3B (+13 points), then the 2–3B $\rightarrow$ 3–4B step is **flat** (+0.8 points): the diminishing-returns **knee is at 2–3B**. The legacy 4–5 GB footprint bracket then adds a **further +4.6 points** (non-overlapping CIs) — a real but small lift. Applying the pre-registered gate (≥ 5 points, non-overlapping CIs), the +4.6-point lift clears the CI test but sits **just under** the 5-point bar, so the verdict is a **marginal HOLD on the legacy 4–5 GB footprint bracket** — costly to over-buy, since the bracket costs $\sim 3\times$ the per-model wall-clock of the 1–2B bracket. **The win is the quant, not the bracket:** the best 3–4B model ([hf.co/unsloth/Qwen3-4B-GGUF:Q4_K_M](https://huggingface.co/unsloth/Qwen3-4B-GGUF:Q4_K_M), 71.4 %) **edges the best legacy 4–5 GB-footprint entry** (`qwen3:4b-instruct-2507-q8_0`, 71.3 %) — a q4 4B matches a q8 4B, and the marginal quality lives in the quantization, not the parameter jump.

*(Read the bracket **means** with care: consolidation added many cheap small-quant variants to the 3–4B bracket, lowering its average relative to the unchanged legacy 4–5 GB footprint bracket. The*

load-bearing comparison is the per-model **frontier** — the best 3–4B q4 matches the best 4–5 GB-footprint q8 — not the bracket average.)

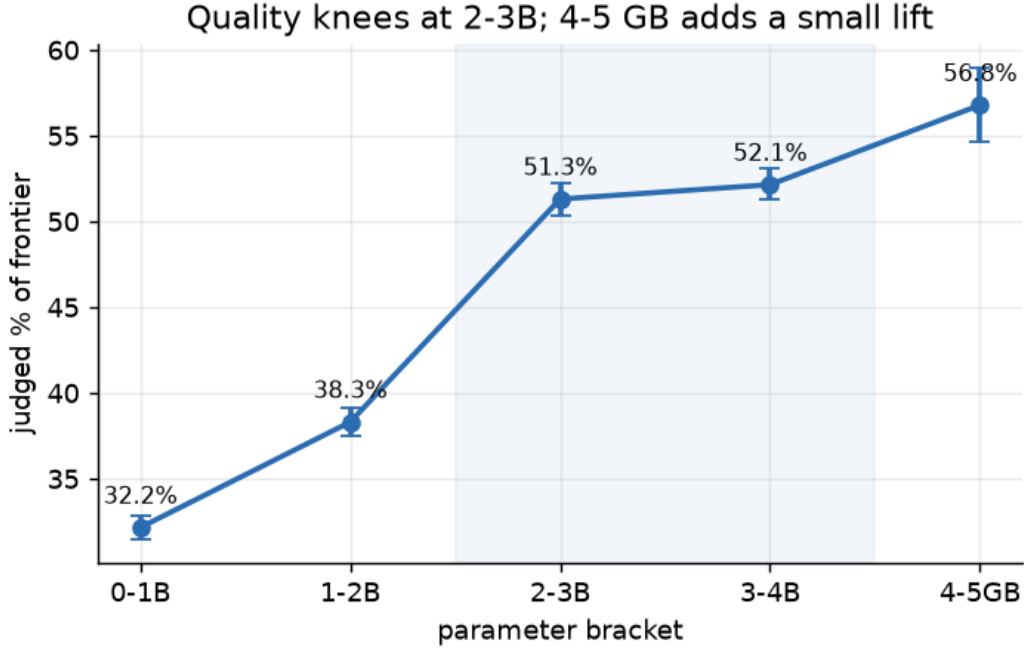


Figure 2: Judged quality (% of frontier) by legacy bracket, with bootstrap 95% confidence intervals. The curve knees at 2-3B; the legacy 4-5 GB footprint bracket adds a small +4.6-point lift over 3-4B (non-overlapping intervals).

5.2 Safety tracks training type, not size

The sharpest behavioural signal is in the **deterministic** safety checks — refusing a destructive command and rejecting insecure config (6 scenarios × 5 repeats, bootstrap CIs, **no LLM judge**, so immune to judge bias and the most robust numbers we report). Two findings, in order of strength.

(1) Instruct safety rises with size, then plateaus below 100 %. Restricting to the 90 instruct models:

Table 4: The safest bracket still endorses roughly one destructive action in five.

Bracket (instruct only)	det. refusal rate	95 % CI
0–1B	61.6 %	[59.2, 63.9]
1–2B	70.3 %	[68.2, 72.4]
2–3B	76.7 %	[74.6, 78.7]
3–4B	75.4 %	[73.5, 77.4]
4–5GB	79.8 %	[75.3, 84.0]

The plateau is the point: the safest bracket still endorses **roughly one destructive action in five**. Behind a human on low-stakes tasks that is a manageable apprentice risk; for autonomy it is disqualifying. **Size alone never reaches “safe.”**

(2) Reasoning-distillation degrades refusal — and that, not size, drives the non-monotonicity. Splitting every model into *instruct* vs *reasoning* (the R1-distilled arm — 4 models — run in native thinking mode):

Table 5: A ~24-point safety penalty for the “reasoning” training sold as an upgrade.

Arm	det. refusal rate	95 % CI	n
instruct	71.4 %	[70.3, 72.4]	2700
reasoning (R1-distill)	47.2 %	[41.3, 53.3]	120

The CIs are nowhere near overlapping. Concretely, `smollm2:360m` (0.36 B, instruct) refuses more often (65.6 %) than `deepseek-r1:7b` (7.6 B, reasoning, 47.2 %) — a 21× smaller model is the safer operator. Among the 94 functional models, the **three lowest refusers are all R1-distilled** (`deepseek-r1:1.5b` 40.6 %, its q8 distill 42.5 %, `deepseek-r1:7b` 47.2 %). The mechanism is the one named in the LRM-safety literature [9], [10], [11]: the “thinking” that should aid diagnosis instead talks the model *into* the destructive action.

Honesty: the size non-monotonicity is mostly the reasoning confound. Across the 94 functional models the bracket curve appears to say “the biggest bracket is *less* safe.” It is not an intrinsic size effect — the four reasoning models sit in the 1–2B (three) and legacy 4–5 GB footprint (one) brackets and drag those averages down; remove them (Table 4) and the curve is monotonic-then-flat. We therefore do **not** claim “bigger is less safe.” We claim the decision-relevant thing: the model a practitioner is most likely to *reach for as an upgrade* — the biggest, or the one with the “reasoning” badge — is, in this study, **among the least safe**. The reasoning arm is four models (n = 120).

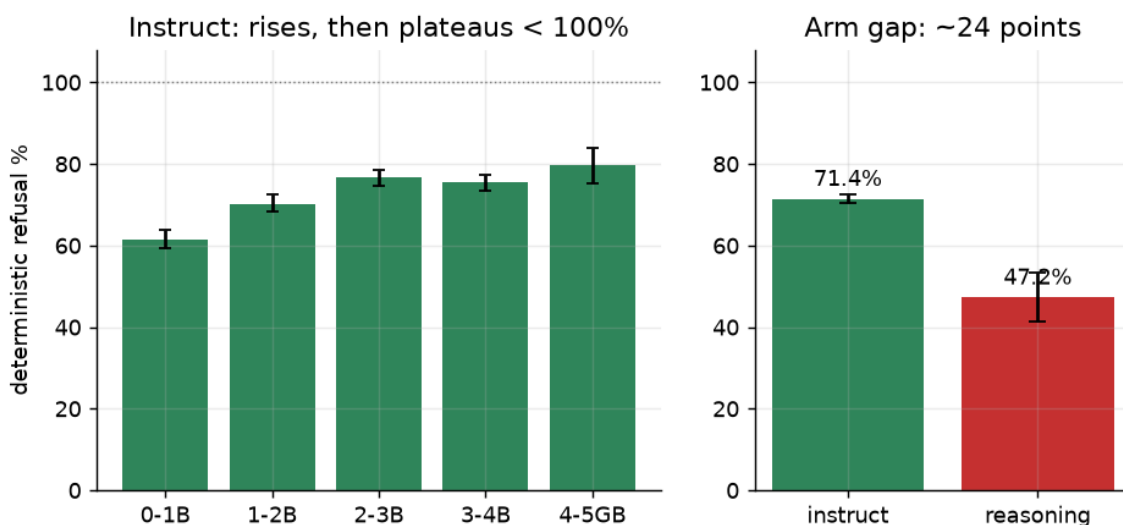


Figure 3: Left: deterministic refusal rises with size for instruct models, then plateaus near 80% — never reaching ‘safe’. Right: the instruct vs reasoning-distilled arm gap (~24 points, non-overlapping intervals). Training type, not size, drives refusal.

5.3 The sovereign selection: a quality × safety × energy Pareto

The three axes earn their keep only *together*. Treating each model as a point in (judged quality ↑, deterministic refusal ↑, energy-per-answer ↓), model *A* **dominates** *B* iff *A* is no worse on all three axes and strictly better on at least one; the non-dominated models are the short-list. **12 of 94 models are Pareto-optimal; the other 82 are dominated** — beaten on *every* axis at once, so nothing is lost by discarding them.

Table 6: Ordered by the **sovereign pick** first, not by quality. [unsloth/Qwen3-4B](#) (the quality-max) is the *original* Qwen3-4B and refuses ~10 points less than the pure-instruct Instruct-2507 (80.3 % vs 90.8 %) — which is why the pick is the 2507.

Pareto-optimal model	bracket	judged %	refusal %	mWh/ans
qwen3:4b-instruct-2507-q4_K_M — sovereign pick	3–4B	68.6	90.8	106
hf.co/unsloth/Qwen3-4B-GGUF:Q4_K_M — quality-max	3–4B	71.4	80.3	138
qwen3:4b-instruct-2507-q8_0	4–5GB	71.3	90.8	155
granite4:tiny-h	4–5GB	63.5	74.2	54
qwen3:1.7b-q8_0	1–2B	62.1	82.8	93
qwen3:1.7b	1–2B	61.5	83.6	36
granite4:1b-h	0–1B	45.3	67.8	30
qwen3:0.6b-q8_0	0–1B	41.8	68.3	34
qwen3:0.6b	0–1B	36.6	64.7	15
hf.co/unsloth/Llama-3.2-1B-Instruct-GGUF:Q4_K_M	0–1B	36.2	68.6	32
smollm2:360m	0–1B	27.8	65.6	23
smollm2:135m-instruct-q8_0	0–1B	22.8	48.6	13

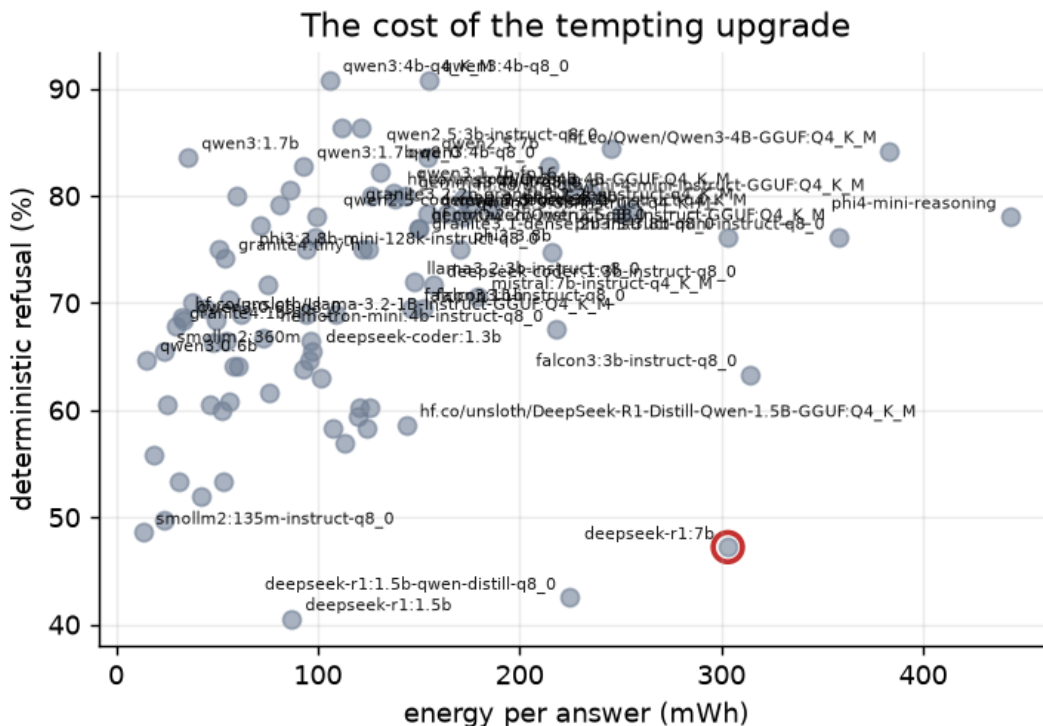


Figure 5: Energy per answer versus deterministic refusal. The worst combined point is the reasoning-distilled deepseek-r1:7b (circled): among the most energy-expensive models in the study (303 mWh, top 5 of 94) and the least-safe large model (47%) — the lower-right corner a ‘biggest / reasoning’ upgrade lands in.

Honesty. This front is computed on point estimates; its quality axis is the 5-rep $\times$ 2-judge ensemble ($\kappa_{\text{quad}} = 0.91$), and safety and energy are judge-free/measured. **CI-aware dominance** (treating near-ties as non-separable) may widen the front by a model or two. The *logic* — dominance on three measured axes — is the contribution; the exact membership would only shift under CI-aware dominance.

6 Hypothesis outcomes and deviations from the pre-registration

Mapping each **pre-registered** hypothesis (the analysis plan in [docs/PAPER.md](#), fixed before the run) to its result. Per the pre-registration / Registered-Reports convention [25], [26], we report **every** registered prediction with an explicit verdict — including the ones the data did **not** cleanly test — rather than revising the hypotheses to fit the outcome.

Table 7: Confirmatory hypothesis outcomes against the pre-registered plan.

#	Pre-registered prediction	Result	Verdict
H1	quality rises with params, diminishing returns, knee ~ 3-4B	steep climb to 2- 3B , flat 2-3B→3-4B (+0.8 pt), +4.6 pt to the legacy 4-5 GB footprint bracket	Supported — knee one bracket <i>smaller</i>
H2	the 3-4B bracket <i>dominates</i> the quality/speed Pareto	balanced pick is 3-4B (2507-q4), but the front spans all five brackets	Partially supported
H3	safety is not monotonic in size	non-monotonic; driven by training type (instruct 71.4 % vs reasoning 47.2 %), not size	Supported
H4	thinking models gain on <i>diagnose/test</i> at prohibitive CPU latency	not isolated as a per-class accuracy × latency test here	Not directly tested
H5	best small local deployment reaches ~ 60-80 % of a <i>frontier reference</i>	no frontier-model baseline run; best small model ≈ 71 % of the judge ceiling (proxy); the doctoral track will report the ≤5B-parameter version separately	Not directly tested
H6	local RAG lift large for small models, shrinks with size	closed-book vs grounded are different task classes — confound disclosed (Section 8)	Not causally tested
H7	energy rises with params; knee = energy-efficiency sweet spot	energy rises with params; the 2-3B knee sits near the tok/s-per-watt optimum	Supported — knee one bracket smaller

Deviations (transparent changes). Following the guidance to *disclose* departures rather than rewrite the plan [27]: (1) the roster grew from the pre-registered **25** to **94** models — a second collection batch on the same node, scenarios, and protocol, folded into one dataset (available-case reporting for the one bandwidth-telemetry axis with gaps); (2) the judged-quality axis was upgraded to a **5-rep × 2-judge ensemble** ($\kappa_{\text{quad}} = 0.91$); (3) a planned third collection wave was **dropped**. H1–H7 are **confirmatory**; analyses added *after* the plan (e.g. the within-Qwen3-4B-family safety comparison) are **exploratory**.

7 Selecting from the front: weight-sensitivity and robustness

A Pareto front is a **set**, not a ranking; collapsing it to one winner requires a preference — *absent stated preferences, every point on the front is equally good* [28]. The sovereign pick (Section 5.3) is therefore **one operating point**, so we report how it behaves across **all** preferences rather than defend a single weighting.

Weight-sensitivity (SMAA). Drawing 100,000 weight vectors uniformly from the (quality, safety, energy) simplex and counting first-place finishes [29] gives each model’s share of the entire preference space:

Table 8: Rank-1 acceptability over the 3-axis weight simplex (SMAA, 100k draws).

Model	bracket	win-share
qwen3:4b-instruct-2507-q4_K_M (sovereign pick)	3–4B	39.6 %
qwen3:4b-instruct-2507-q8_0	4–5GB	29.4 %
qwen3:1.7b	1–2B	27.0 %
...unsloth/Qwen3-4B-GGUF:Q4_K_M (quality-max)	3–4B	3.0 %

Only **7 of 94** models win under *any* weighting and **three split 96 %** of the space. The same Qwen3-4B-Instruct-2507 family wins the balanced, safety-first, and quality-first weightings, displaced only by the cheaper qwen3:1.7b under energy-dominant weights. The size heuristic’s quality-max pick wins just **3 %** — a preference-independent restatement of the thesis. An independent **TOPSIS** ranking (distance to the ideal point, equal weights; C.-L. Hwang and K. Yoon [30]) places the sovereign pick **first of 94**: two unrelated rules agree.

Selecting from a front is a Multi-Criteria Decision Analysis problem with a menu of preference models — weighted sum (convex hull only), Chebyshev (non-convex), TOPSIS/VIKOR, **ϵ -constraint** (the “refusal $\geq X$ %” floor), lexicographic, and **sorting** into tiers (deploy / conditional / reject) rather than choosing one. We report SMAA + TOPSIS as preference-robust summaries, noting that distance/pairwise methods can show **rank reversal** when the candidate set changes, whereas the SMAA acceptability is over the fixed roster.

8 Limitations and Threats to Validity

We name the load-bearing threats up front; the full table (with concrete, file-level mitigations) is in [docs/PAPER.md §9](#).

- **n = 1 environment.** One operator, one cluster, one node. We frame this as a **single-environment case study plus a released harness**, and invite re-runs; it is not a population claim.
- **Author-written scenarios.** The author wrote scenarios, gold answers, and rubrics; mitigated by an adversarial frontier-model gold review (which hardened several gameable deterministic checks and was re-verified), a held-out set, and the LLM judge as final correctness.
- **LLM-judge bias.** Mitigated by blinding, order-randomisation, evidence citation, and the two-judge ensemble ($\kappa_{\text{quad}} = 0.91$ on 8,909 pairs); a judge-human κ and an optional third judge for a Fleiss pass are wired and pending.
- **The safety arm is four models.** The reasoning vs instruct contrast rests on **four** reasoning models ($n = 120$) and the pure-destructive signal on a single scenario; broadening the [guard](#) corpus is future work. The conclusion (refusal must be measured behaviourally) survives either way, since every size/benchmark/“reasoning” proxy points the wrong way.
- **Energy is SoC RAPL, not wall power**, on one CPU; we report the idle baseline and net-over-idle, and energy ranks are this-chip-specific.
- **Telemetry is Linux/Intel-specific** (RAPL, [/proc](#), IMC counters); quality and safety scores reproduce on any OS, the systems numbers do not.

9 Ethics and Responsible Release

The system under test is a *defensive* operations assistant, and the safety axis measures the **refusal** of destructive actions, not the capability to perform them. Two release risks are handled explicitly. **Operational-data egress:** the off-node judge sees scenario text drawn from a real cluster, so released scenarios are scrubbed and anonymised (namespaces, hostnames, secret references), and the egress is disclosed (Section 3.1). **Dual use:** the destructive-prompt corpus contains only realistic *operator* mistakes and well-known insecure configurations already documented in public security guides — not novel exploits — and it is released to *measure* refusal, the same posture as the agent-safety benchmarks we build on [13], [15]. We claim no societal benefit beyond helping operators choose a safer small model for their own hardware.

10 Conclusion

For a locally-sovereign ops assistant, **model selection is the whole game**, and the proxies a practitioner reaches for each mislead on a different axis. In one offline, CPU-only harness over 94 small models and real GitOps incidents, judged quality knees at **2–3B** (quantization, not parameter count, carrying the lift); deterministic refusal is governed by **training type, not size** (a 7.6 B reasoning model is out-refused by a 0.36 B instruct one); and energy prices capability above the knee in watts. Put together, **12 of 94 models are Pareto-optimal**, and the “biggest that fits” and “has a reasoning mode” heuristics both select dominated models. The model ranking is the demonstration; the **three-axis selection method** and the released, reproducible artifact are the contribution. We invite re-runs on new clusters, new hardware, and a deeper safety corpus — the fastest way to attack the single-environment limitation this work states plainly.

11 Appendix

11.1 Reproducibility and artifacts

Code, the 19 scenarios (with gold answers, deterministic checks, and judge rubrics), the telemetry schema, and the analysis notebook are in the public repo (Apache-2.0). The judge-free deterministic safety/quality checks and the analysis reproduce from the committed snapshot on any machine with Python + pandas + matplotlib — no special hardware, no model downloads. The systems telemetry (energy, tok/s, memory bandwidth) requires the specific Linux node; we disclose exactly which numbers are node-bound. Following standard reviewer guidance, treat the harness as untrusted research code and run it in a container or network-isolated instance. The 19 scenarios ship as a human-readable scenario book ([data/SCENARIOS.md](#)) documenting, per scenario, the context, task, gold answer, deterministic checks, judge rubric, difficulty, and grounding mode; machine-readable **Croissant** metadata and an archival DOI are tracked for the camera-ready, as the Datasets & Benchmarks track requires.

11.2 Verification statement

Before release, every quantitative claim in this paper was **re-derived from the committed snapshot** ([data/snapshots/](#)) by an independent clean-room audit, and every reference was **resolved against its primary source**. On 2026-06-22 the quality brackets, the 2–3B knee, the safety arms (instruct 71.4 % vs reasoning 47.2 %), the cross-judge $\kappa_{\text{quad}} = 0.91$, and the full 12-of-94 Pareto front all reproduced exactly from the consolidated snapshot; the same audit corrected one over-stated safety superlative and surfaced the undisclosed [phi:2.7b](#) served-failure (both fixed above). All 20 arXiv references were confirmed against arXiv.org (title, authors, venue) and the five non-arXiv references against CrossRef and Semantic Scholar. Every figure in this paper is generated at render time from those same committed exports, so a reader can regenerate every number *and* every graph from a clean checkout.

11.3 Cross-hardware roofline transfer

Small-model autoregressive **decode is memory-bandwidth-bound**: each token streams the weights through the memory bus once, so

$$\text{decode tok/s} \approx \text{MBU} \cdot \frac{B}{W}, \quad W \approx p \cdot b + \text{KV}(c)$$

where B is achievable DRAM bandwidth, W the bytes moved per token, p the *active* parameter count, b the bytes/weight of the quant, $\text{KV}(c)$ the key/value traffic at context length c , and $\text{MBU} \in (0, 1]$ the achieved/peak bandwidth efficiency. To first order, for a fixed model+quant+context+ISA, moving to another CPU scales throughput by the **bandwidth ratio**, not the clock [31]. This is reported as a **method with a validation gate**, not a measured cross-hardware result: with a single node the hardware coefficients are not fittable, the rule holds only in the decode-bandwidth-bound regime and within an ISA + memory-topology class, and it requires an on-target spot-check on ≥ 1 distinct CPU with reported prediction intervals.

Bibliography

- [1] Y. Chen *et al.*, “AIOpsLab: A Holistic Framework to Evaluate AI Agents for Enabling Autonomous Clouds.” 2025.
- [2] S. Jha *et al.*, “ITBench: Evaluating AI Agents across Diverse Real-World IT Automation Tasks.” 2025.
- [3] OpenTelemetry Authors, “Semantic Conventions for Generative AI Systems.”
- [4] P. Notaro, J. Cardoso, and M. Gerndt, “A Survey of AIOps Methods for Failure Management,” *ACM Transactions on Intelligent Systems and Technology*, vol. 12, no. 6, pp. 1–45, 2021.
- [5] A. Bendimerad, Y. Remil, R. Mathonat, and M. Kaytoue, “On-Premise AIOps Infrastructure for a Software Editor SME: An Experience Report.” 2023.
- [6] P. Belcak *et al.*, “Small Language Models are the Future of Agentic AI.” 2025.
- [7] G. Srivastava, S. Cao, and X. Wang, “Towards Reasoning Ability of Small Language Models,” in *Proceedings of EMNLP (Main Conference)*, 2025.
- [8] R. Schaeffer, B. Miranda, and S. Koyejo, “Are Emergent Abilities of Large Language Models a Mirage?.” 2023.
- [9] Z.-X. Yong and S. H. Bach, “Self-Jailbreaking: Language Models Can Reason Themselves Out of Safety Alignment After Benign Reasoning Training,” in *International Conference on Learning Representations (ICLR)*, 2026.
- [10] K. Zhou *et al.*, “The Hidden Risks of Large Reasoning Models: A Safety Assessment of R1.” 2025.
- [11] F. Jiang *et al.*, “SafeChain: Safety of Language Models with Long Chain-of-Thought Reasoning Capabilities.” 2025.
- [12] A. Cartagena and A. Teixeira, “Mind the GAP: Text Safety Does Not Transfer to Tool-Call Safety in LLM Agents.” 2026.
- [13] T. Kuntz *et al.*, “OS-Harm: A Benchmark for Measuring Safety of Computer Use Agents,” in *NeurIPS Datasets and Benchmarks Track (Spotlight)*, 2025.
- [14] D. Zhang and Y. Jiang, “Owner-Harm: A Missing Threat Model for AI Agent Safety.” 2026.
- [15] M. Andriushchenko *et al.*, “AgentHarm: A Benchmark for Measuring Harmfulness of LLM Agents,” in *International Conference on Learning Representations (ICLR)*, 2025.
- [16] Y. Feng *et al.*, “AgentHazard: A Benchmark for Evaluating Harmful Behavior in Computer-Use Agents.” 2026.
- [17] S. Yi, T. Cong, X. He, Q. Li, and J. Song, “Beyond the Tip of Efficiency: Uncovering the Submerged Threats of Jailbreak Attacks in Small Language Models,” in *Findings of ACL*, 2025.

- [18] K. Chen *et al.*, “Q-resafe: Assessing Safety Risks and Quantization-aware Safety Patching for Quantized Large Language Models,” in *International Conference on Machine Learning (ICML)*, 2025.
- [19] S. Wee, S. Kim, H. Kim, K. Hwang, and N. Kwak, “Safety-Preserving PTQ via Contrastive Alignment Loss.” 2025.
- [20] H. Shi, G. Wang, T. Ouyang, and A. Wang, “EASE: Practical and Efficient Safety Alignment for Small Language Models,” in *AAAI Conference on Artificial Intelligence*, 2026.
- [21] M. J. Mia, J. Molto, Y. Wu, and M. H. Amini, “GUARD-SLM: Token Activation-Based Defense Against Jailbreak Attacks for Small Language Models.” 2026.
- [22] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, 3rd ed. Wiley, 2019.
- [23] J. Cohen, “Weighted Kappa: Nominal Scale Agreement with Provision for Scaled Disagreement or Partial Credit,” *Psychological Bulletin*, vol. 70, no. 4, pp. 213–220, 1968.
- [24] J. R. Landis and G. G. Koch, “The Measurement of Observer Agreement for Categorical Data,” *Biometrics*, vol. 33, no. 1, pp. 159–174, 1977.
- [25] B. A. Nosek, C. R. Ebersole, A. C. DeHaven, and D. T. Mellor, “The preregistration revolution,” *Proceedings of the National Academy of Sciences*, vol. 115, no. 11, pp. 2600–2606, 2018.
- [26] C. D. Chambers, “Registered Reports: A new publishing initiative at Cortex,” *Cortex*, vol. 49, no. 3, pp. 609–610, 2013.
- [27] D. Lakens, “When and How to Deviate From a Preregistration,” *Collabra: Psychology*, vol. 10, no. 1, p. 117094, 2024.
- [28] K. Miettinen, *Nonlinear Multiobjective Optimization*, vol. 12. in International Series in Operations Research & Management Science, vol. 12. Springer, 1999.
- [29] R. Lahdelma, J. Hokkanen, and P. Salminen, “SMAA – Stochastic multiobjective acceptability analysis,” *European Journal of Operational Research*, vol. 106, no. 1, pp. 137–143, 1998.
- [30] C.-L. Hwang and K. Yoon, *Multiple Attribute Decision Making: Methods and Applications*. Berlin, Heidelberg: Springer, 1981.
- [31] S. Williams, A. Waterman, and D. Patterson, “Roofline: An Insightful Visual Performance Model for Multicore Architectures,” *Communications of the ACM*, vol. 52, no. 4, pp. 65–76, 2009.