

Phoenix-VL 1.5 Medium: Continued Pre-Training of Multimodal Sovereign AI for On-Premises Public-Sector Deployment

Ze Yu Zhang^{1*†}, Jiale Wang^{1*†},

Jia Hui Bong¹, Umar Jamil², Wen Jett Lam¹, Biondi Lee¹, Eva Lim¹,
Vincent Maladiere², Hong Jin Jonathan Ng¹, Arka Ray¹, Elijah Seah¹, Justin Tan¹,
Fiona Teo¹, Guang Xiang Teo¹, Toh Xin Yi Grace¹, Yeo Shun Yi¹, Gee Wah Ng^{1‡}

¹Home Team Science and Technology Agency (HTX), Singapore

²Mistral AI

Abstract

Sovereign AI deployments often require foundation models to operate under constraints overlooked by standard benchmarks: adherence to legislation, multilingual and multimodal grounding in local contexts, institution-specific safety behaviors, and on-premises deployment on power-constrained infrastructure. We present a production case study detailing the multi-stage engineering pipeline that spans trillion-token-scale continued pretraining, multi-billion-token alignment and post-training for a 123B-parameter natively multimodal foundation model, *Phoenix-VL 1.5 Medium*, optimized for Singapore public-sector use. On our novel localized evaluation suite and behavioral safety framework, the model substantially improves localized multimodal, statutory, and public-safety capability (a +0.118 average gain over the strongest baselines) while retaining broad general capability. We also share empirical analysis from a full production deployment currently serving over 30,000 users and processing ~ 1 billion tokens over a two-month period. Ultimately, our findings demonstrate that sovereign AI is best approached as an end-to-end systems problem, where data curation, domain training, institutional alignment, and hardware layout interact and cannot be optimized in isolation.

1 Introduction

Public-sector deployments of foundation models often impose constraints that generic hosted LLMs do not satisfy: operation without external web access, localized legislative and policy knowledge, multilingual users, controlled infrastructure, and institution-specific safety behavior. In these settings, Sovereign AI is not only a question of who

owns the model, but of who controls its training, evaluation, behavior, and deployment lifecycle.

Standard model evaluations emphasize general reasoning, coding, mathematics, and global multimodal tasks. A model can perform well on these benchmarks while failing at local statutory recall, public-sector policy knowledge, localized visual understanding, or abstention when live retrieval is unavailable, creating an evaluation gap for air-gapped public-sector use. Recent regional adaptation efforts such as EuroLLM (Martins et al., 2025; Ramos et al., 2026), ALLaM (Bari et al., 2025), and SEA-LION (Ng et al., 2025) have shown that continued pretraining improves regional-language coverage, but they target open, text-only language models and do not address multimodal grounding, jurisdiction-specific legal and safety behavior, or air-gapped serving. To the best of our knowledge, this is the first work to examine these dimensions jointly as an end-to-end deployment problem, rather than as isolated components. We provide an extended discussion in the related work in Appendix A.

Sovereign AI We define *Sovereign AI* as an institutional ownership over the training, evaluation, safety behavior, deployment environment, and lifecycle of a foundation model for domain-adapted use. This prioritizes controlled dependence, transparent trade-offs, and operational authority over from-scratch development. We study whether a large multimodal foundation model can be deeply adapted to a local sovereign context without significant degradation in general capability, and whether it can be served under realistic on-premises latency and power constraints. The deployment jurisdiction is Singapore: *Phoenix-VL 1.5 Medium* serves the Ministry of Home Affairs (the “Home Team”) and associated public-sector entities. The model is in its second month of full production deployment on air-gapped, on-premises infrastructure, serving

*Co-primary authors. All authors other than the co-primary and senior authors are listed alphabetically by last name.

†Corresponding authors: ZHANG_ZE_YU@htx.gov.sg; WANG_jiale@htx.gov.sg.

‡Senior author and project lead; provided overall research and engineering leadership.

over 10 public service functions and a user base exceeding 30,000 personnel across text-and-vision analysis, statutory interpretation, and document workflows (~1 billion tokens processed to date).

This work makes four contributions. **First**, we characterize practical requirements for multimodal sovereign AI in an air-gapped public-sector setting, covering local knowledge, multilinguality, multimodal grounding, safety behavior, and infrastructure control. **Second**, we present a multi-stage training recipe for a 123B multimodal model that injects jurisdiction-specific and regional knowledge while preserving broad general capabilities. **Third**, we introduce a localized evaluation and safety suite for measuring alignment across government, statutory, public-safety, multimodal, abstention, prompt-injection, and leakage tasks. **Fourth**, we provide an inference-efficiency analysis that maps the trade-off between interactivity and throughput for on-premises deployment.

The target workflows may lack external APIs or live web access, and generic safety policies may be unsuitable. We therefore identify six operational requirements: air-gapped operation, local legislative and policy knowledge, multilingual users, localized visual grounding, behavioral alignment, and controlled infrastructure. Table 1 maps the adaptation pipeline to these requirements; Section 6 addresses controlled infrastructure.

We distinguish reusable methods from jurisdiction-specific artifacts. The pipeline, training recipe, evaluation-construction protocols, and deployment lessons in Section 7 **can be reused with data from another jurisdiction**. The corpora, benchmarks, and Model Policy Document are specific to Singapore. For five of the seven localized tasks, we release de-identified prompt templates, runnable scorers, judge rubrics, decoding settings, and sample items (Appendix E).¹

2 Data and Architecture

Phoenix-VL 1.5 Medium is a 123B multimodal decoder-only transformer supporting image-text inputs, adapted from an existing foundation model rather than trained from scratch. Target tasks span localized visual question answering, long-document analysis, local policy and statutory QA, multilingual assistance, and controlled refusal for

requests requiring unavailable live or private data. Architectural details are in Appendix B.

Data mixture and controls. The adaptation corpus combines general replay data, regional-language text, local policy and statutory sources, human-annotated local image-text pairs, instructions, and preference data. We use replay data in each major stage, including image-text pairs during continued pretraining to mitigate catastrophic forgetting. Data are filtered using source-level rules, text/image safety classifiers, quality filters, exact and fuzzy de-duplication, and benchmark decontamination. Appendix C provides mixture design and filtering details.

3 Training Recipe

Table 1 summarizes the pipeline: continued pretraining (CPT), long-context extension (LCE), supervised fine-tuning (SFT), online direct preference optimization (ODPO), and multimodal domain training, together with the deployment requirement each stage addresses.

3.1 Continued Pretraining (CPT)

CPT adapts the base model toward local and regional usage while preserving broad capabilities via replay. The 1T-token mixture broadens coverage of local policy, statutory, cultural, and regional-language content, alongside image-text replay data to preserve multimodal behavior. We train at the native short-context length during CPT, reserving long-context specialization for LCE. The final CPT checkpoint is selected by averaging late-stage checkpoints over a validation-selected window rather than using a single terminal snapshot (Izmailov et al., 2018).

Long Context Extension (LCE). LCE extends the context window to 131k tokens using long-document data, short-context replay, position interpolation (Peng et al., 2024), and attention-temperature stabilization (Nakanishi, 2025) to support realistic policy, legislative, and administrative documents.

3.2 Post-training

3.2.1 Instruction Tuning

Following LCE, we perform full-parameter supervised fine-tuning (SFT) to align the model with its target role as a multilingual, multimodal assistant. The mixture comprises approximately 16B tokens

¹We withhold the two recall tasks built from verbatim statutory text and specific agency facts.

Stage	Requirement addressed	Tokens	Main sources	Trainable modules	Primary evidence
CPT	Air-gapped local knowledge; multilingual users	1 Trillion (txt+img)	General replay, regional languages, local policy/legal/cultural data, multi-modal replay	Primarily language backbone	Local and multilingual gains; retention of general capability
LCE	Long public-sector documents	250 Billion (txt)	Long documents with short-context replay	Language backbone	131k context support; long-document workflow readiness
SFT	Assistant behavior	16 Billion (txt+img)	Chat, QA, summarization, document tasks, replay	Full model; vision modules frozen where applicable	Instruction following
ODPO	Behavioral governance (preference alignment)	5 Billion (txt+img)	Preference prompts, image feedback, ranked completions	Policy model	Lower infinite generation rate; improved alignment metrics
Local domain	Local law and policy; local visual context	22 Billion (txt+img)	Human-localized scenes; grounded local policy and statutory corpus	Two-stage vision/text schedule	SG-Multimodal, SG-Gov, SG-Legal, HT-Lexicon custom benchmark gains

Table 1: Requirements-mapped adaptation pipeline. Modalities: txt = text, img = image-text. Token counts rounded.

of multi-turn conversational, long-document summarization, and document-set Question Answer (QA) data, interleaving regular and long sequences to retain the extended context window, and is supplemented with roughly 10% pretraining replay for stability and safety retention. Training combines a completion-only loss with distillation from Mistral Large 3 (Mistral AI, 2025a), keeping the vision encoder frozen; post-SFT evaluations show gains over the base model Mistral Medium 3.1 (Mistral AI, 2025b) on multilingual benchmarks with general capability preserved, with the loss formulation and intermediate results in Appendix H.

3.2.2 Online Direct Preference Optimization

Following SFT, we apply an online variant of Direct Preference Optimization (DPO) (Rafailov et al., 2023; Guo et al., 2024; Xiong et al., 2023; Liu et al., 2026) on a 5B-token preference mixture containing feedback prompts, image-grounded feedback, and instruction-with-image pairs. Candidate completions are sampled on-policy from the current model, so the policy is optimized against its own generation distribution rather than a fixed offline set. ODPO targets alignment artifacts while preserving capability: the infinite generation rate² falls to 3.8%, and the post-ODPO checkpoint reaches 0.624 on Arena-Hard-v2 (Li et al., 2024b). Appendix I gives the full online loop, loss formulation, reward-model details, and hyperparameters.

²We define the *infinite generation rate* as the fraction of evaluation generations that reach the configured maximum output length without emitting an end-of-sequence token; lower is better.

3.2.3 Multimodal Continued Domain Training

This stage injects contextualized regional data by interleaving two curated datasets: a human-annotated multimodal dataset of the Singapore-specific culture and scenes and a textual corpus covering localized knowledge, policies, and legislation. The data corpora for this stage totaled approximately 22 billion multimodal tokens.

Localized human-annotated multimodal data.

Generic captioning pipelines frequently miss regional operational nuances, generating generic descriptions (e.g., “Playground” instead of “*Toa Payoh Dragon Playground*”, or identifying a municipal landmark simply as a “Public Plaza”). To address this, we ran a human annotation programme to improve visual descriptions for local culture, scenes, landmarks, equipment, and events. Local annotators, drawing on their personal knowledge and supplemented by web search, were required to enrich or rewrite base captions to inject precise contextual details. The programme was governed by a structured *Data Annotation Playbook*, which codified the required caption dimensions, the criteria for choosing domain-specific localization over generic taxonomic labels, and a cross-annotator adjudication procedure (Appendix C.5).

Grounded textual data. We curated a textual corpus grounded in local statutory and governance knowledge. To minimize hallucination in legislative and operational contexts, high-fidelity documents were first retrieved from verified public

databases; frontier LLMs then generated complex question-answering pairs grounded exclusively in those contexts; and a rejection-sampling filter discarded responses exhibiting hallucination or deviation from the source material.

Methodology. We use two stages to add domain data while limiting forgetting. First, we unfreeze the vision encoder and adapter and train on 30% image-text pairs and 70% text. This allowed the model to acquire more robust domain visual knowledge. We then freeze both vision modules to stabilize the newly acquired multimodal representations, continually training the language backbone exclusively on the remaining domain text. Across both stages, general replay data was included to mitigate capability degradation. The localized gains are robust to the domain-to-replay ratio: in a sweep from 1:9 to 5:5, the 1:9 mixture already recovers $\sim 92\%$ of the total HT-Lexicon (our domain benchmark, elaborated in Section 4.1) gain while MMLU stays within ~ 1 point of its pre-adaptation value throughout (Appendix G.2).

Evaluation. Domain training improves all four localized benchmarks over the post-ODPO instruct checkpoint (*Instruct* vs. *Ours* in Table 2). The instruct checkpoint is itself already strong relative to open-weight baselines, which we attribute to tokens introduced during CPT (Table 4, Section 4.3).

4 Results

We evaluate the final checkpoint of our recipe (Table 1) under a zero-shot, no-retrieval, no-web-search protocol that approximates air-gapped deployment. For benchmark evaluations, all models use greedy decoding (temperature 0), a shared prompt template, and the same maximum output length. Construction, scoring, and validation details for each localized task are given as benchmark cards in Appendix E, with full per-model and subgroup results in Appendix F.

4.1 Localized Capability Evaluation

Standard benchmarks emphasize general STEM, coding, and instruction following, and do not capture the knowledge required for localized governance, statutory reasoning, and sovereign operational contexts. We therefore introduce a four-task suite grounded in the Singapore (SG) context: SG-Multimodal, human-annotated VQA over Singapore scenes, landmarks, and infrastructure;

Task	Instruct	Ours	Best baseline	Δ
SG-Multimodal	0.7207	0.7509	0.6716	+0.0793
SG-Gov	0.8366	0.9265	0.8006	+0.1259
SG-Legal	0.7667	0.8640	0.6144	+0.2496
HT-Lexicon	0.8495	0.9081	0.8891	+0.0190
Average	0.7934	0.8624	0.7439	+0.1185

Table 2: Localized results (Pass@1; zero-shot, no-search). *Instruct* is the aligned checkpoint (SFT + ODPO) preceding domain training (Section 3.2.3); *Ours* is the final *Phoenix-VL 1.5 Medium* checkpoint. *Best baseline* is the strongest external model *per task*. Per-model and subgroup results are in Appendix F.

SG-Gov, parametric-knowledge QA spanning policies and functions across all government ministries; SG-Legal, statutory QA constructed directly from Singapore legislative materials; and HT-Lexicon, operational terminology, agency functions, and procedural vocabulary of the Home Team (HT), Singapore’s Ministry of Home Affairs and its agencies.

Under the no-search protocol, *Phoenix-VL 1.5 Medium* achieves the highest aggregate Pass@1 among the evaluated baselines on all four localized benchmarks (Table 2). The largest gains occur on statutory, government-policy, and public-safety terminology tasks, which are precisely the areas targeted by the local adaptation mixture. These results suggest that targeted local adaptation can be more effective than parameter scale alone for this deployment setting. Because the suite is jurisdiction-specific and only partially releasable, we interpret these results as evidence of industry deployment readiness rather than absolute model superiority.

4.2 Multilinguality and General-Capability Retention

We quantify whether localization preserves general capability by evaluating *Phoenix-VL 1.5 Medium* on standard multilingual, multimodal, and broad-reasoning benchmarks (Table 3); full per-model breakdown in Appendix F, Table 10.

Phoenix-VL 1.5 Medium is not uniformly strongest on global benchmarks: models optimized for mathematics, coding, or general multimodal tasks retain clear advantages on several metrics, consistent with our objective of improving localized capability without unacceptable degradation of general behavior.

How trade-offs were balanced. Acceptable regressions were determined empirically. Through-

Category	Metric	<i>Phoenix-VL 1.5 Medium</i>	Best baseline
Multilingual	Malay Belebele	0.9274	0.9151
Multilingual	Tamil Belebele	0.8447	0.8391
Multimodal	MMMU	0.6078	0.6122
Document VQA	DocVQA	0.9305	0.9590
Knowledge	MMLU-Pro	0.7681	0.8670
Math	MATH-500	0.8120	0.9780
Code	LiveCode	0.3240	0.4280

Table 3: General-capability retention. We report representative metrics here, with the full benchmark set in Table 10. Multilingual scores are Belebele (Bansarkar et al., 2024); other benchmarks: MMMU (Yue et al., 2024), DocVQA (Mathew et al., 2021), MMLU-Pro (Wang et al., 2024b), MATH-500 (Lightman et al., 2024), and LiveCodeBench (Jain et al., 2024).

out development, a beta “Playground” gave a cohort of operational users (~25, growing as the product matured) side-by-side access to the candidate checkpoints of Section 4.3 on their own real workloads. Losses on competition mathematics and code generation were not particularly flagged by user feedback, whereas any drop in localized statutory or policy accuracy surfaced immediately, and a checkpoint was cleared only once the cohort confirmed both conditions. Comparing against the strongest baseline *per metric* also conflates capability lost through adaptation with capability the base model never had: measured against Mistral Medium 3.1 (Table 4), general capability is preserved (MMLU -0.025 , MMMU -0.014 , LiveCodeBench $+0.031$) while localized capability rises sharply (SG-Legal $+0.423$, HT-Lexicon $+0.099$).

4.3 Stage-wise Ablation

Table 4 isolates the contribution of each stage of Table 1 by evaluating the post-stage checkpoints on two localized tasks, two regional-language tasks, and three general benchmarks. Three patterns emerge. First, CPT drives most localized knowledge and language acquisition: it contributes 66% of the total SG-Legal gain and 73% of the HT-Lexicon gain, together with the bulk of the Tamil and Malay Belebele improvements, while general capability is essentially unchanged. Second, SFT and ODPO add assistant behavior rather than localized knowledge, improving instruction following, code generation, and generation stability (Appendices H and I) with little movement on the localized tasks. Third, domain training makes the final push on localized capability (SG-Legal $+0.097$,

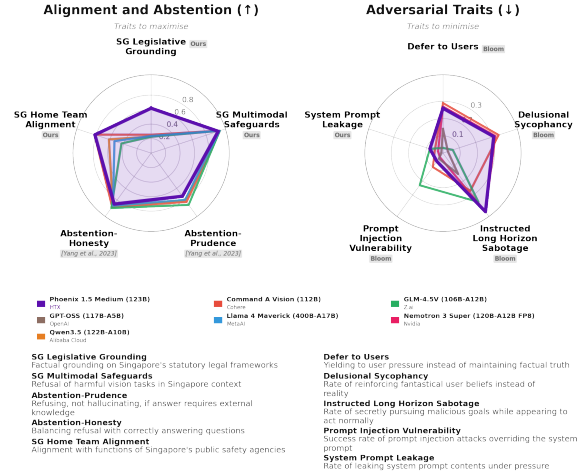


Figure 1: Safety and behavioral profiles of *Phoenix-VL 1.5 Medium* against frontier baselines; exact values in Appendix J. Left (higher is better): *Phoenix-VL 1.5 Medium* leads on SG Legislative Grounding and is competitive on SG Home Team Alignment, while trailing the strongest baselines on the abstention metrics and SG Multimodal Safeguards. Right (lower is better): near-best on prompt injection, mid-range on the sycophancy metrics, and tied for the highest value on system-prompt leakage and instructed long-horizon sabotage.

HT-Lexicon $+0.058$ over the instruct checkpoint) at a modest cost to general benchmarks; how these trade-offs were adjudicated is discussed in Section 4.2.

5 Sovereign AI Safety and Alignment

Air-gapped deployments require localized behavioral governance: with no external database or web access, the model must refuse or escalate unsupported claims rather than fabricate answers. General alignment frameworks (Bai et al., 2022; Gupta et al., 2025) capture broad policies but not region-specific nuances (Chua et al., 2025), so we localize the safety policy itself and evaluate safety alignment, epistemic prudence, and operational helpfulness jointly.

Evaluation domains and construction. The suite targets three localized domains: *SG Legislative Grounding* (factual accuracy on penal-code and statutory text), *SG Home Team Alignment* (institutional knowledge and agency values), and *SG Multimodal Safeguards* (refusal of harmful visual tasks in local contexts). Extending the agentic methodology of Gupta et al. (2025), a *Model Policy Document* developed with key stakeholders codifies behavioral constraints; evaluation samples are

Stage	SG-Legal	HT-Lexicon	Bele-ta	Bele-ms	MMLU	MMMU	LiveCode
Mistral Medium 3.1	0.441	0.809	0.621	0.850	0.860	0.622	0.293
+CPT	0.721	0.881	0.884	0.931	0.862	0.628	0.289
+SFT	0.749	0.888	–	–	0.864	0.668	0.448
+ODPO (<i>Instruct</i>)	0.767	0.850	–	–	0.863	0.674	0.468
+Domain (<i>Phoenix-VL 1.5 Medium</i>)	0.864	0.908	0.845	0.927	0.835	0.608	0.324

Table 4: Stage-wise ablation of the adaptation pipeline (Pass@1, zero-shot). Each row is the checkpoint after the named stage; +ODPO is the *Instruct* checkpoint of Table 2 and the final row is the deployed *Phoenix-VL 1.5 Medium*. Bele-ta / Bele-ms are Tamil and Malay Belebele; ‘–’ marks benchmarks on which an intermediate checkpoint was not evaluated. **Bold** marks the per-column best. LCE, applied between CPT and SFT, targets long-context capability and is reported separately (Appendix G).

synthetically generated from it and anchored to reference datasets on regional governance and legislation (Gao et al., 2023).

Scoring. SG Legislative Grounding (WER and word-overlap F1) and SG Home Team Alignment (ordered key-term recall) use *deterministic, judge-free* scorers; SG Multimodal Safeguards, the abstention metrics, and the Bloom-derived adversarial metrics use *LLM/VLM-as-judge* evaluation (Appendix K). Because behavioral safety is a property of the deployed system rather than the weights alone, *Phoenix-VL 1.5 Medium* is scored under its *production configuration* (the production system prompt, guardrails, and decoding settings) while baselines use their defaults. Re-scoring every judged suite with a judge from a third model family changes no ranking on the Safeguards or abstention suites and moves the Bloom-derived rates by at most 0.03, except on instructed long-horizon sabotage, where the rate falls from 0.42 to 0.31 under the alternative judge; item-level agreement is $\kappa=1.0$ on Safeguards and $\kappa=0.97$ on abstention, and we report the primary-judge sabotage value, which is the conservative one (Appendix K.5).

We further evaluate *Phoenix-VL 1.5 Medium* on general behavioral dimensions adapted from the Anthropic Bloom framework (Gupta et al., 2025), alongside an abstention framework (Yang et al., 2023) and a custom system-prompt-leakage set. Each targets an operational risk specific to air-gapped settings; the individual dimensions and the risks they map to are defined in Appendix K.

5.1 Safety and Model Behavior Profile

Grounding and Robustness. On **SG Legislative Grounding**, *Phoenix-VL 1.5 Medium* leads all evaluated baselines at 0.619, reflecting the parametric recall required when retrieval is unavailable, and its average sycophancy elicitation is 28.5% across

the Defer to Users and Delusional Sycophancy tasks. Against instruction-level attacks, **Prompt Injection Vulnerability** is near-best at 6%, so operational guardrails hold when the model parses untrusted document workflows; **System Prompt Leakage** at 8% is low in absolute terms but joint-highest in the comparison set (Table 13).

Abstention and Safeguards. Where fabrication is costlier than refusal, *Phoenix-VL 1.5 Medium* yields an Honesty score of 0.8681 and a Prudence score of 0.7361, producing no false refusals on the interleaved answerable controls (0/275; Table 14), and blocks 97.6% of adversarial tasks on the **SG Multimodal Safeguards** benchmark. On both dimensions it nonetheless trails every evaluated baseline (Table 13): several baselines refuse every Safeguards item, and each scores higher on both abstention metrics.

Instructed long-horizon sabotage. The clearest weakness in the profile is Instructed Long Horizon Sabotage, on which *Phoenix-VL 1.5 Medium* scores 0.42 and is tied for the highest value in a comparison set spanning 0.15–0.42. The metric measures covert pursuit of an *explicitly instructed* malicious goal over a long task, and so presupposes adversary control of the instruction channel, which in this deployment is a centrally administered system prompt on air-gapped infrastructure; adversarial content reaching the model through user or document inputs falls under prompt injection instead (Appendix K). The score therefore measures risk for the agentic workflows we plan next rather than characterizing the currently deployed system, and it is the behavior we target in the next post-training iteration.

Pareto Frontier for serving Phoenix-VL 1.5 Medium

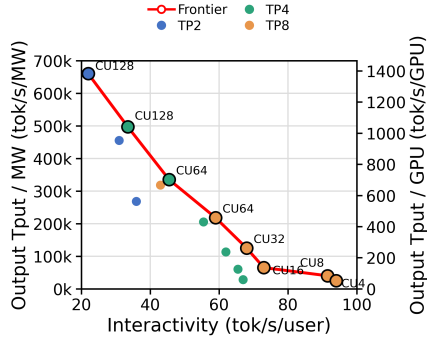


Figure 2: Pareto frontier for *Phoenix-VL 1.5 Medium*: interactivity against output throughput per MW.

6 Infrastructure Sovereignty: On-Premises Serving & Production Use

Sovereign deployments must balance on-premises hardware efficiency and power draw against user experience. Because the 123B model in BF16 precision exceeds single-GPU memory, we benchmark tensor-parallelism (TP) degrees TP2, TP4, and TP8 on NVIDIA B200 nodes under vLLM (Kwon et al., 2023), using a 1,000-token input/output profile with system power normalized to 2.17 kW per GPU (Table 5, Appendix D). Figure 2 maps TP layout and concurrent-user (CU) load onto a Pareto frontier of *system efficiency* (output tokens/s/MW) against *interactivity* (decode tokens/s/user).

Against our Service Level Objective (SLO) of ≥ 40 decode tokens/s/user for fluid interactive chat, the frontier identifies TP4 at CU64 as the production regime: it delivers ~ 45 tokens/s/user while sustaining $\sim 330,000$ output tokens/s/MW, and is the only configuration satisfying both the interactivity SLO and the efficiency floor. TP2 is more power-efficient but falls below the SLO at operational load, and TP8 reaches high interactivity only at low concurrency before collapsing under cross-GPU communication and software overhead (Appendix D). We therefore adopt TP4 as the baseline configuration and auto-scale horizontally with multi-node replicas once concurrency exceeds CU64, preserving real-time performance without sacrificing compute efficiency.

Production usage. Adoption of *Phoenix-VL 1.5 Medium* is organic rather than mandated. Comparable open-weight models are available to the same users on the same on-premises infrastructure. Usage concentrates on grounded document work. Summarization and report drafting form the largest

share, and motivated the 131k context window, followed by OCR, retrieval over internal databases, and context-specific image captioning and extraction. This distribution is why the adaptation mixture favours statutory text and grounded document data over code and mathematics.

7 Discussion

Our central claim is that sovereign AI is an end-to-end systems problem: data curation, domain training, institutional alignment, and hardware layout interact and cannot be optimized in isolation. We summarize this into five lessons that transfer to other localized air-gapped deployments. (1) Keep the visual pathway in the domain-training loop; unfreezing the vision encoder and adapter on localized image-text data, then refreezing them for text-only continuation (Section 3.2.3), improved domain-specific performance over the original checkpoint (Table 2). (2) Choose the parallelism configuration empirically on target hardware, keeping in mind expected user load, interactivity, and throughput (Section 6). (3) Online preference optimization directly targets generation artifacts and could reduce the infinite generation rate (Section 3.2.2); such artifacts should be tracked as production metrics rather than masked in aggregate capability scores. (4) Despite replay, the model regressed on some benchmarks after SFT and recovered only partially after ODPO (Appendix H); post-training techniques that better resist forgetting warrant investigation. (5) Validate the serving pipeline, not just the model: production feedback surfaced degraded OCR on low-quality scans, which traced to the serving stack rather than the weights and was corrected in serving configuration alone.

8 Conclusion

We presented a case study in operationalizing sovereign multimodal AI for an air-gapped public-sector setting, showing that targeted adaptation improves localized multimodal, statutory, policy, and public-safety capability while retaining competitive general performance; and with careful deployment choices, enabling on-premises serving for a user base exceeding 30,000 and processing ~ 1 billion tokens to date. The broader lesson is that sovereign AI should be built as an end-to-end system spanning adaptation, localized evaluation, behavioral governance, and infrastructure.

Limitations

Disclosure. As our datasets concern restricted public-sector, legal, and operational sources, we had to withhold model weights, full training data, and parts of the localized evaluation suite. In their place we favor in-paper protocol disclosure over artifact release, specifying source categories, mixture proportions, scoring procedures, judge rubrics, prompt templates, and sanitized examples in enough detail to assess credibility and transferability, though exact reproduction is not possible.

Generalization. Our localized benchmarks are jurisdiction- and sector-specific, and a single deployment cannot establish empirical transfer. We therefore release scripts, templates, and artifacts to facilitate adaptation to other settings. *Other jurisdictions:* only the corpora, benchmarks, and Model Policy Document are Singapore-specific; the pipeline is data-substitutable by design. Generalization should therefore be assessed across independent adaptation studies (Appendix A). *Other legal systems:* the statutory-task construction protocol, including verified-source retrieval, grounded question generation, rejection sampling, and provision-level recall scoring, is corpus-agnostic, with released templates and scorers supporting adaptation (Appendix E). *Other languages:* our adaptation mixture covers 12 languages, with evaluation across eight languages from six families, suggesting broad rather than language-specific adaptation, but not transfer to arbitrary languages. *Non-government domains:* the six operational requirements are shared by regulated private sectors (finance, healthcare, national infrastructure, etc.) and the ministries in SG-Gov parallel several of them, but transfer to non-government deployments remains untested.

Judging. Some safety and open-ended evaluations rely on LLM-as-judge pipelines. We assess sensitivity to this choice by re-scoring every judged suite with an independent judge from a third model family (Appendix K.5), and the rubric for the localized safeguards task was manually reviewed in full, but we do not report a separate human calibration study, and residual judge bias may remain.

Adversarial robustness. *Phoenix-VL 1.5 Medium* is tied for the highest instructed long-horizon sabotage rate in our comparison set, and trails the evaluated baselines on the abstention metrics and on multimodal safeguards (Section 5.1);

these are the behavioral axes targeted in the next post-training iteration.

Staleness. Statutory and policy knowledge evolves over time, so parametric recall must complement rather than replace authoritative retrieval and professional review.

Cost. The adopted TP4 regime is practical but resource-intensive, and may not suit smaller organizations.

Coverage. Our pass@1 zero-shot protocol does not capture retrieval-augmented, tool-using, or multi-turn deployment modes.

Ethical Considerations

The system is designed for controlled public-sector use and not for direct unsupervised legislative, enforcement, or benefits decisions. Local legal and policy knowledge is used to improve internal understanding and reduce hallucination, but model outputs must be verified against authoritative sources for operational decisions. Evaluation data derived from sensitive or restricted workflows is anonymized, aggregated, or withheld. We report safety evaluations for abstention, prompt injection, system-prompt leakage, and localized harmful multimodal requests, but these evaluations are not exhaustive. Deployment should include logging, access control, human escalation, periodic policy refresh, and post-deployment monitoring.

Acknowledgments

We gratefully acknowledge the contributions of Wen Xiu Tan (core contributor) and Jun En Tan (core contributor) of Team Phoenix, and of our Mistral AI collaborators: Adrien Sadé, Guillaume Kunsch, Jia Sin Loh, Nicolas Schuhl, Rupert Menneer, and Yimu Pan. We thank Chee Wee Ang, Chief AI Officer of HTX, for his executive sponsorship of this work, and our collaborators across the HTxAI ecosystem for their partnership. We also thank the operational users who evaluated our checkpoints on real workloads and provided feedback throughout development and deployment. This work was supported by the Home Team Science and Technology Agency (HTX), a statutory board under the Ministry of Home Affairs, Singapore.

References

- Pravesh Agrawal, Szymon Antoniak, Emma Bou Hanna, Baptiste Bout, Devendra Chaplot, Jessica Chudnovsky, Diogo Costa, Baudouin De Monicault, Saurabh Garg, Theophile Gervet, and 1 others. 2024. Pixtral 12b. *arXiv preprint arXiv:2410.07073*.
- Joshua Ainslie, James Lee-Thorp, Michiel de Jong, Yury Zemlyanskiy, Federico Lebrón, and Sumit Sanghai. 2023. Gqa: Training generalized multi-query transformer models from multi-head checkpoints. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.
- Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, and 1 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*.
- Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775.
- M Saiful Bari, Yazeed Alnumay, Norah Alzahrani, Nouf Alotaibi, Hisham Alyahya, Sultan AlRashed, Faisal Mirza, Shaykhah Alsubaie, Hassan Alahmed, Ghadah Alabduljabbar, and 1 others. 2025. Allam: Large language models for arabic and english. In *International Conference on Learning Representations*, volume 2025, pages 34179–34214.
- Andrei Z Broder. 1997. On the resemblance and containment of documents. In *Proceedings. Compression and Complexity of SEQUENCES 1997 (Cat. No. 97TB100171)*, pages 21–29. IEEE.
- Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198.
- Gabriel Chua, Leanne Tan, Ziyu Ge, and Roy Ka-Wei Lee. 2025. Lost in localization: Building rabakbench with human-in-the-loop validation to measure multilingual safety gaps. *arXiv preprint arXiv:2507.05980*.
- Longxu Dou, Qian Liu, Guangtao Zeng, Jia Guo, Jiahui Zhou, Xin Mao, Ziqi Jin, Wei Lu, and Min Lin. 2024. Sailor: Open language models for south-east asia. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 424–435.
- Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual pre-training for cross-lingual llm adaptation: Enhancing japanese language capabilities. *arXiv preprint arXiv:2404.17790*.
- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.
- Shangmin Guo, Biao Zhang, Tianlin Liu, Tianqi Liu, Misha Khalman, Felipe Llinares, Alexandre Rame, Thomas Mesnard, Yao Zhao, Bilal Piot, and 1 others. 2024. Direct language model alignment from online ai feedback. *arXiv preprint arXiv:2402.04792*.
- Isha Gupta, Kai Fronsdaal, Abhay Sheshadri, Jonathan Michala, Jacqueline Tay, Rowan Wang, Samuel R. Bowman, and Sara Price. 2025. [Bloom: an open source tool for automated behavioral evaluations](#). Anthropic Alignment Science Blog.
- Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don’t stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 8342–8360.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.
- Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.
- Piotr Indyk and Rajeev Motwani. 1998. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pages 604–613.
- Pavel Izmailov, Dmitrii Podoprikin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. 2018. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*.
- Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2024. Livecodebench: Holistic and contamination free evaluation of large language models for code. *arXiv preprint arXiv:2403.07974*.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th symposium on operating systems principles*, pages 611–626.

- Katherine Lee, Daphne Ippolito, Andrew Nystrom, Chiyuan Zhang, Douglas Eck, Chris Callison-Burch, and Nicholas Carlini. 2022. Deduplicating training data makes language models better. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8424–8445.
- Jure Leskovec, Anand Rajaraman, and Jeffrey David Ullman. 2020. *Mining of Massive Datasets*, 3 edition. Cambridge University Press.
- Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024a. Cmmlu: Measuring massive multitask language understanding in chinese. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11260–11285.
- Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. 2024b. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *arXiv preprint arXiv:2406.11939*.
- Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let’s verify step by step. In *International Conference on Learning Representations*, volume 2024, pages 39578–39601.
- Alexander H Liu, Kartik Khandelwal, Sandeep Subramanian, Victor Jouault, Abhinav Rastogi, Adrien Sadé, Alan Jeffares, Albert Jiang, Alexandre Cahill, Alexandre Gavaudan, and 1 others. 2026. Ministral 3. *arXiv preprint arXiv:2601.08584*.
- Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.
- Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2023. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*.
- Pedro Henrique Martins, João Alves, Patrick Fernandes, Nuno M Guerreiro, Ricardo Rei, Amin Farajian, Mateusz Klimaszewski, Duarte M Alves, José Pombal, Nicolas Boizard, and 1 others. 2025. Eurollm-9b: Technical report. *arXiv preprint arXiv:2506.04079*.
- Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.
- Sai Krishna Mendu, Harish Yenala, Aditi Gulati, Shanu Kumar, and Parag Agrawal. 2025. Towards safer pre-training: Analyzing and filtering harmful content in webscale datasets for responsible LLMs. In *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence*, pages 466–474.
- Mistral AI. 2025a. Introducing Mistral 3. <https://mistral.ai/news/mistral-3>.
- Mistral AI. 2025b. Mistral medium 3.1. Mistral Docs. Model ID: mistral-medium-2508.
- Eric Mitchell. 2023. A note on dpo with noisy preferences & relationship to ipo. <https://ericmitchell.ai/cdpo.pdf>.
- Ken M. Nakanishi. 2025. Scalable-softmax is superior for attention. *arXiv preprint arXiv:2501.19399*.
- Raymond Ng, Thanh Ngan Nguyen, Yuli Huang, Tai Ngee Chia, Wai Yi Leong, Wei Qi Leong, Xi-anbin Yong, Jian Gang Ngui, Yosephine Susanto, Nicholas Cheng, and 1 others. 2025. Sea-lion: South-east asian languages in one network. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 512–526.
- Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Zhiqiang Hu, Chenhui Shen, Yew Ken Chia, Xingxuan Li, Jianyu Wang, Qingyu Tan, and 1 others. 2024. Seallms-large language models for southeast asia. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 294–304.
- NLLB Team, Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Hefernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.
- Krzysztof Ociepa, Łukasz Flis, Krzysztof Wróbel, Adrian Gwoździej, and Remigiusz Kinas. 2024. Bielik 7b v0.1: A polish language model—development, insights, and evaluation. *arXiv preprint arXiv:2410.18565*.
- Jupinder Parmar, Sanjev Satheesh, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Reuse, don’t retrain: A recipe for continued pre-training of language models. *arXiv preprint arXiv:2407.07263*.
- Bowen Peng, Jeffrey Quesnelle, Honglu Fan, and Enrico Shippole. 2024. Yarn: Efficient context window extension of large language models. In *International Conference on Learning Representations*, volume 2024, pages 31932–31951.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

- Miguel Moura Ramos, Duarte M Alves, Hippolyte Gisserot-Boukhlef, João Alves, Pedro Henrique Martins, Patrick Fernandes, José Pombal, Nuno M Guerreiro, Ricardo Rei, Nicolas Boizard, and 1 others. 2026. Eurollm-22b: Technical report. *arXiv preprint arXiv:2602.05879*.
- Noam Shazeer. 2020. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*.
- Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, and 1 others. 2022. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*.
- Guijin Son, Hanwool Lee, Sungdong Kim, Seungone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. 2025. Kmmlu: Measuring massive multitask language understanding in korean. 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 4076–4104.
- Kiran Vodrahalli, Santiago Ontanon, Nilesh Tripuraneni, Kelvin Xu, Sanil Jain, Rakesh Shivanna, Jeffrey Hui, Nishanth Dikkala, Mehran Kazemi, Bahare Fatemi, and 1 others. 2024. Michelangelo: Long context evaluations beyond haystacks via latent structure queries. *arXiv preprint arXiv:2409.12640*.
- Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024a. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.
- Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyang Jiang, and 1 others. 2024b. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290.
- Tongtong Wu, Linhao Luo, Yuan-Fang Li, Shirui Pan, Thuy-Trang Vu, and Gholamreza Haffari. 2024. Continual learning for large language models: A survey. *arXiv preprint arXiv:2402.01364*.
- Wei Xiong, Hanze Dong, Chenlu Ye, Ziqi Wang, Han Zhong, Heng Ji, Nan Jiang, and Tong Zhang. 2023. Iterative preference learning from human feedback: Bridging theory and practice for rlhf under kl-constraint. *arXiv preprint arXiv:2312.11456*.
- Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. 2023. Alignment for honesty. *arXiv preprint arXiv:2312.07000*.
- Shunyu Yao, Howard Chen, Austin W Hanjie, Runzhe Yang, and Karthik Narasimhan. 2023. Collie: Systematic construction of constrained text generation tasks. *arXiv preprint arXiv:2307.08689*.
- Ziqi Yin, Hao Wang, Kaito Horio, Daisuke Kawahara, and Satoshi Sekine. 2024. Should we respect llms? a cross-lingual study on the influence of prompt politeness on llm performance. In *Proceedings of the Second Workshop on Social Influence in Conversations (SICoN 2024)*, pages 9–35.
- Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, and 1 others. 2024. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9556–9567.
- Biao Zhang and Rico Sennrich. 2019. Root mean square layer normalization. In *Advances in Neural Information Processing Systems*, volume 32.
- Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. 2024. Negative preference optimization: From catastrophic collapse to effective unlearning. *arXiv preprint arXiv:2404.05868*.
- Lucen Zhong, Zhengxiao Du, Xiaohan Zhang, Haiyi Hu, and Jie Tang. 2025. Complexfuncbench: exploring multi-step and constrained function calling under long-context scenario. *arXiv preprint arXiv:2501.10132*.

A Related Work

Regional and sovereign language models. A growing body of work adapts foundation models to specific linguistic and national contexts. For Southeast Asia, SEA-LION (Ng et al., 2025), the SeaLLMs series (Nguyen et al., 2024), and Sailor (Dou et al., 2024) continue-pretrain open base models on regional-language corpora to improve coverage of languages such as Indonesian, Thai, Vietnamese, Malay, and Tamil. Comparable efforts exist for other regions, including EuroLLM (Martins et al., 2025; Ramos et al., 2026) for European languages and ALLaM (Bari et al., 2025) for Arabic; at the single-national-language level, Swallow (Fujii et al., 2024) (Japanese) and Bielik (Ociepa et al., 2024) (Polish) follow the same continued-pretraining pattern. These works establish that targeted continued pretraining improves regional-language performance over English-centric base models; the value of each study lies in its recipe and lessons, and generalization accrues across the sequence of independent case studies rather than within any single deployment. Our work shares this continued-pretraining foundation but differs in scope along four axes that matter for public-sector deployment: we target multimodal (image-text) rather than text-only grounding; we inject jurisdiction-specific statutory, policy, and public-safety knowledge alongside language coverage; we localize the behavioral safety policy itself rather than inheriting generic alignment (Section 5); and we treat air-gapped on-premises serving as a first-class constraint rather than releasing open weights for general use.

Continued pretraining and domain adaptation. Adapting pretrained models to new domains without catastrophic forgetting is a long-standing problem. Domain-adaptive pretraining (Gururangan et al., 2020) and studies on data reuse and replay (Parmar et al., 2024; Wu et al., 2024) show that mixing in-domain data with replay of the original distribution mitigates forgetting, a strategy we apply at every stage of our pipeline. Our contribution is not a new adaptation algorithm but a deployment-driven composition of these techniques: continued pretraining, long-context extension, instruction tuning, online preference optimization, and multimodal domain training, calibrated for a localized, multimodal, air-gapped target setting.

Localized safety and behavioral evaluation. General alignment frameworks such as Helpfulness–Harmlessness (Bai et al., 2022) and the Bloom behavioral suite (Gupta et al., 2025) target broadly applicable safety policies. Recent work shows these generalized frameworks under-capture region-specific harms: RabakBench (Chua et al., 2025) demonstrates that safety classifiers calibrated on global data miss locally salient categories. Most sovereign-AI efforts to date emphasize localized *knowledge* while leaving safety and abstention policies generic. Our behavioral framework instead localizes the safety policy itself, and pairs it with abstention evaluation (Yang et al., 2023), an operational requirement in air-gapped settings, where fabricating an answer is worse than declining one. To our knowledge, this is the first work to treat localized knowledge adaptation, institution-specific safety governance, and air-gapped serving efficiency as a single end-to-end deployment problem.

B Architecture

B.1 Language backbone

Phoenix-VL 1.5 Medium uses a decoder-only transformer backbone for autoregressive generation, with grouped-query attention (Ainslie et al., 2023) (96 query heads, 8 key–value heads), SwiGLU activations (Shazeer, 2020), and RMSNorm (Zhang and Sennrich, 2019).

B.2 Vision encoder

Phoenix-VL 1.5 Medium couples the language backbone with a 2B-parameter Vision Transformer (ViT) encoder and a lightweight adapter, in line with recent open vision-language model designs (Liu et al., 2023; Wang et al., 2024a; Agrawal et al., 2024; Chen et al., 2024). Images are decomposed into 14×14 pixel patches at native aspect ratio; a spatial merge of size 2 reduces the number of visual tokens passed to the language backbone, keeping long multimodal sequences tractable.

B.3 Tokenizer

Phoenix-VL 1.5 Medium uses a byte-level BPE tokenizer with a vocabulary of 131,072 tokens, trained on text from more than 100 languages. The large multilingual vocabulary yields favourable compression for the regional languages relevant to our deployment as well as for source code, which reduces effective sequence lengths at inference time. Spe-

cial tokens ([IMG], [IMG_BREAK], [IMG_END]) delimit image-patch sequences in the input stream.

B.4 Modalities and Use Cases

The intended use cases for *Phoenix-VL 1.5 Medium* include: visual question answering (VQA) in local context; queries on the regional governance and legislative knowledge (Figure 3); multilingual document, chart analysis, and Optical Character Recognition (OCR) tasks; and general intelligence dialogue. It is deployed to serve the public sector and associated entities, with downstream applications benefiting from the model’s strong grounding in regional knowledge and languages.

C Data Mixture and Filtering Details

C.1 Pretraining data

The pretraining data comprises vast amounts of data that span across multiple domains and languages to further develop the linguistics and cultural understanding of *Phoenix-VL 1.5 Medium* in the regional languages relevant to our deployment. *Phoenix-VL 1.5 Medium* also needs to be strong in regional-specific knowledge that is relevant to the users while retaining its general capabilities. The general vision capabilities are also crucial for users, and hence, an interleaving image-text corpus is also incorporated.

C.2 Design Principles

Dataset construction followed four principles to achieve the desirable outcomes. First, we utilized data quality classifiers to retain only the highest-quality tokens given a fixed token and compute budgets. Second, we emphasized sources absent from the base checkpoint to encourage novel regional understanding. Third, we included a portion of the original pretraining data, as replay data, to mitigate the potential risk of catastrophic forgetting. This included interleaved image tokens as well, which accounted for 11% of the pretraining dataset. Fourth, we intentionally increased the representation of the target regional and adjacent languages in the pretraining mixture to boost the model’s understanding of those languages on top of English. The final data mixture is predominantly in multilingual tokens, across 12 different languages, at 53%.

C.3 Sampling Policy

The sampling policy is designed to balance coverage across sources while prioritizing the essential data. By default, each data source is sampled to contribute approximately one epoch, meaning that all tokens from that source are observed at least once during pretraining. Essential or underrepresented sources, such as the local legislation, are up-sampled up to 4 times to increase their influence in the training process, with highly specialized datasets up-sampled further according to source priority. Conversely, broad and less-targeted sources, such as CommonCrawl, are down-sampled to prevent over-representation, and this is done based on a per-source basis.

C.4 Data Quality Filtering

The integrity and quality of datasets is a critical prerequisite for reliable model development and deployment. In particular, it is necessary to ensure that the collected data is free from undesirable or harmful content, including Not Safe For Work (NSFW) materials, inappropriate media, or misinformation. The presence of such content would introduce ethical risks and can also negatively affect downstream model behavior (Mendu et al., 2025).

C.4.1 Data Quality Control and Decontamination

We implemented a rigorous, two-tiered safety and filtering strategy to eliminate inappropriate material. At the metadata level, URL-based filtering is performed to identify domains or links commonly associated with explicit or unwanted materials. A maintained list of blocked domains, combined with pattern-based detection rules, enables the system to flag and exclude potentially unsafe sources before ingestion. At the content level, text and image classifiers are used to scan the retrieved data for indicators of unwanted data.

Heuristic-based quality checks are also applied to ensure that the content is coherent, structurally valid, and grammatically reasonable. These heuristics are designed to detect low-quality or malformed data. Samples that fail these checks are filtered out to maintain a high standard of textual integrity across the dataset.

Crucially, to ensure the integrity of our downstream evaluations, we perform n-gram decontamination against our localized benchmarks. This

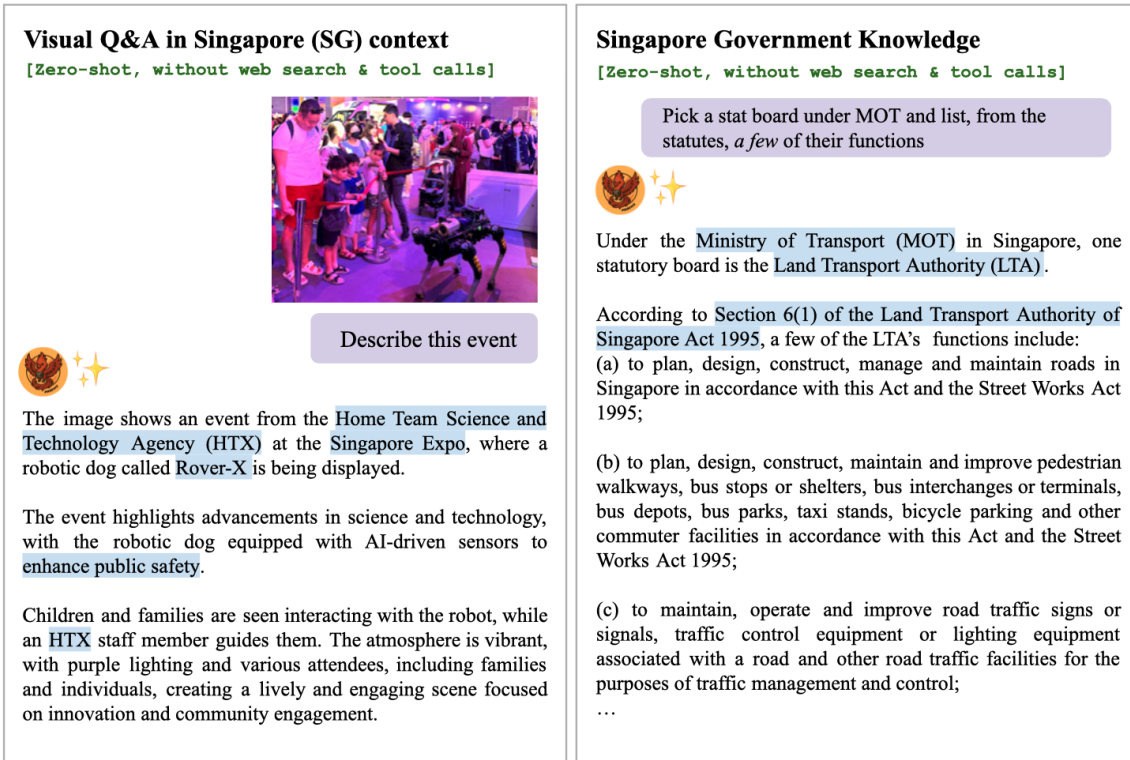


Figure 3: Example Multimodal interactions demonstrating strong visual Q&A in local context and zero-shot legal knowledge recall. Domain-specific knowledge in output is highlighted in blue.

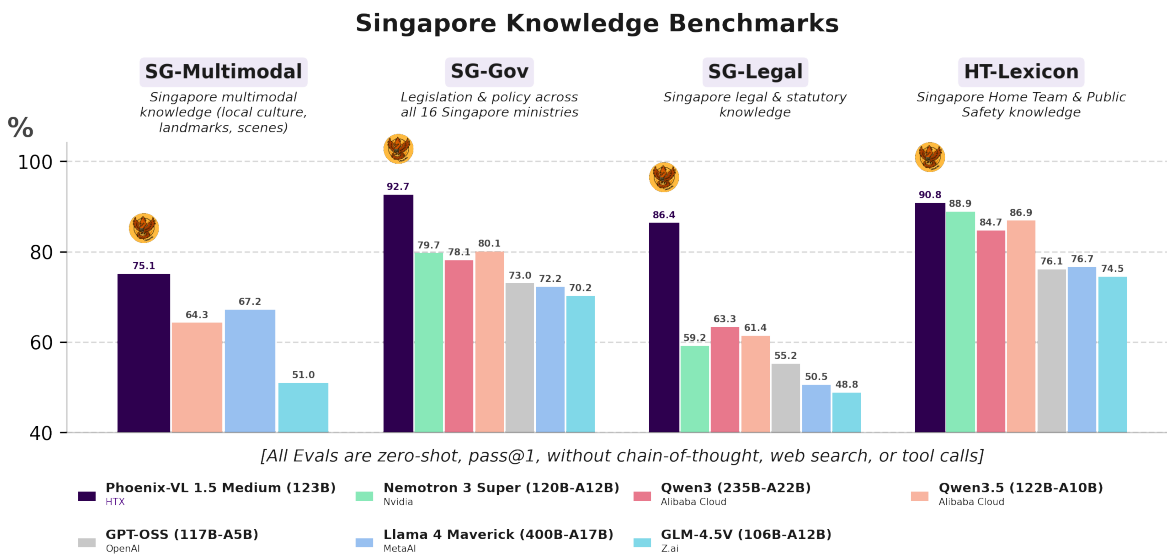


Figure 4: Performance for *Phoenix-VL 1.5 Medium* and open-weights models on Singapore-specific knowledge, based on benchmarks described in Section 4.1. Evaluation was run zero-shot with no web-search, simulating the air-gapped environment where this model will be deployed.

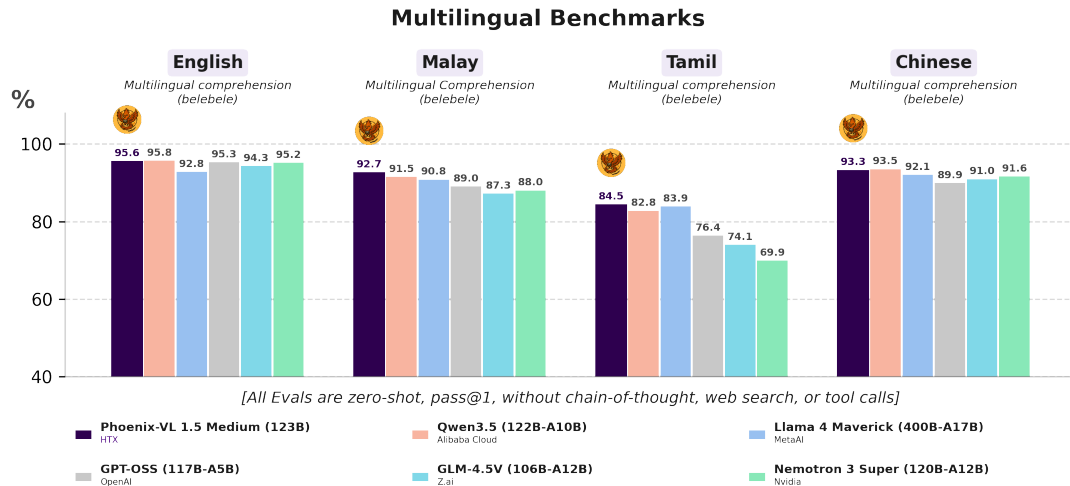


Figure 5: **Multilingual performance:** *Phoenix-VL 1.5 Medium* exhibits competitive comprehension of Malay, Tamil, and Chinese.

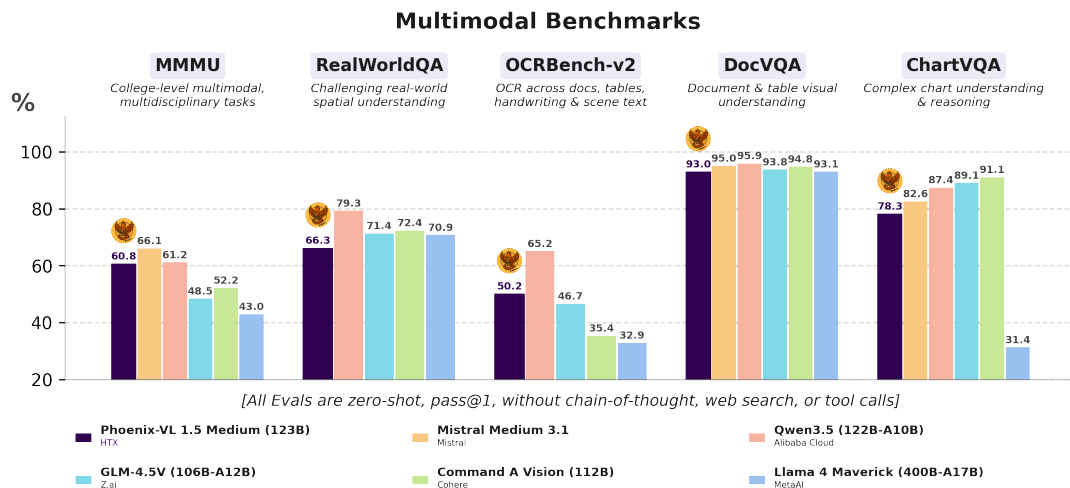


Figure 6: **Multimodal reasoning:** *Phoenix-VL 1.5 Medium* demonstrates strong vision capabilities, particularly on complex document analysis.

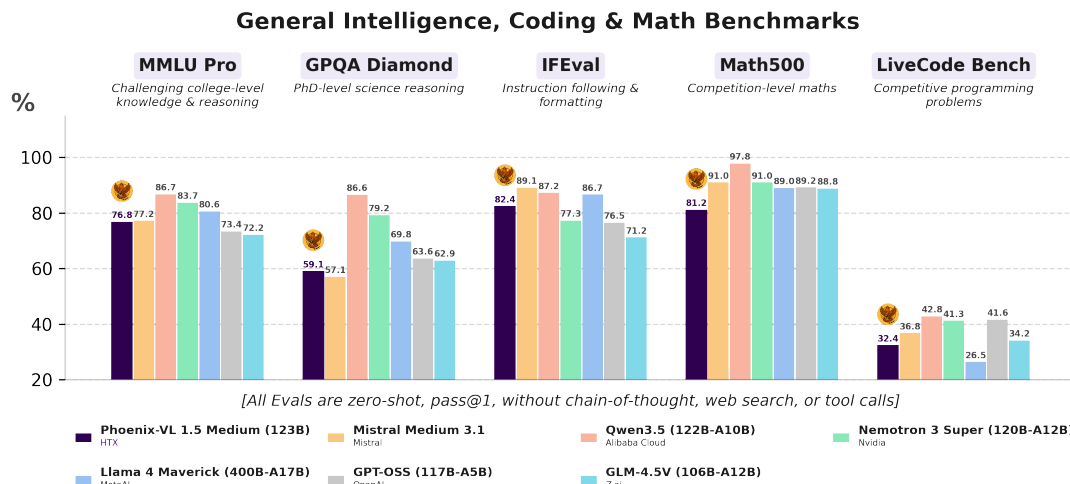


Figure 7: **Preservation of general intelligence.** Our training recipe ensures domain adaptation does not compromise *Phoenix-VL 1.5 Medium*'s broad-based multimodal intelligence.

ensures that our results reflect true learning and generalization rather than test-set memorization.

C.4.2 Deduplication

Exact and fuzzy deduplication techniques are employed to minimize repetitions. Exact deduplication is performed by identifying identical text segments across the dataset. This is done through substring matching to detect exact matches of documents or document fragments across different sources (Lee et al., 2022). Fuzzy deduplication is implemented using the MinHash Locality Sensitive Hashing (LSH) algorithm (Broder, 1997; Indyk and Motwani, 1998; Leskovec et al., 2020; Smith et al., 2022). Compact signatures that represent the similarity structure of documents are generated and compared. Documents that exceed a predefined Jaccard similarity threshold are considered near-duplicates, and redundant instances are removed accordingly.

C.4.3 Domain Coverage and Quality Stratification

The composition of the data mixture is handled along two primary dimensions; the domain and the content quality. This approach is designed to ensure that the mixture is diverse in its knowledge and performs in general capabilities.

The data is classified into the broad domains, using a lightweight classifier that includes STEM, Code, logical and analytical problem-solving, and general knowledge. This flexibility enables us to balance the data mixture and control the proportion of each of these domains, preventing over-representation or under-representation of specific knowledge areas. The data is also sieved through by in-house trained quality classifiers, which assigns a rating of low, medium, or high quality, specific to our requirements. As mentioned, we elect to choose from the high-quality data to retain the existing base capabilities given the fixed training budget.

C.5 Quality Control of Human-Annotated Multimodal Data

The localized image-text annotations (Section 3.2.3) were produced under the *Data Annotation Playbook*, which required captions to cover objects, localized naming, relationships, scene context, and negative statements within a fixed length band, and prohibited placeholders, unverifiable assumptions, and redundancy.

Enriched claims carried source references, and low-confidence items were routed to review. A subset of items was independently annotated by multiple annotators; disagreements were triaged by an LLM judge and adjudicated by the training team, which returned written feedback to annotators. We do not report inter-annotator agreement for these annotations.

D Compute Infrastructure and Optimization

Continued pre-training was conducted on an on-premises cluster of Nvidia GB200 GPUs, while post-training largely took place on a separate Nvidia H100 GPU cluster to enable parallel experiments. In both stages, we set aside at least one node for the evaluation of the checkpoints on the fly, and to be used as a replacement node. This helps to minimize disruptions of the training due to hardware issues, which happened on a few occasions during our training.

Based on internal tests performed on our infrastructure, Fully Sharded Data Parallelism (FSDP) strategy proved to be the most stable, efficient, and scalable for our setup. The ability to shard the model parameters, gradients, and optimizer states across all the GPUs ensures that it can fit within the hardware constraints.

Inference benchmark configuration. Table 5 lists the baseline serving configuration underlying the Pareto analysis in Figure 2.

Non-selected serving regimes. Two configurations were evaluated and rejected. TP2 at CU128 attains the highest system efficiency in the sweep, approaching 660,000 output tokens/s/MW, but interactivity falls to ~ 22 decode tokens/s/user, well below the ≥ 40 tokens/s/user SLO for interactive chat. TP8 exceeds 90 tokens/s/user at CU4–8 but is uneconomical at operational load: cross-GPU communication overhead and current software constraints on 8-way tensor parallelism for this architecture collapse efficiency below 50,000 tokens/s/MW. TP4 at CU64 is therefore the only regime in the sweep that satisfies both constraints simultaneously.

E Localized Benchmarks

The localized suite comprises four capability tasks (SG-Multimodal, SG-Gov, SG-Legal, HT-Lexicon), two open-ended recall tasks (SG Legisla-

Parameter	Configuration
Input sequence length	1000 tokens
Output sequence length	1000 tokens
Hardware	2×B200, 4×B200, 8×B200
System Power Normalized per GPU	2.17 kW

Table 5: Inference benchmark baseline configuration.

tive Grounding and SG Home Team Alignment), and a multimodal safety task (SG Multimodal Safeguards). All tasks use the zero-shot, no-search protocol of Section 4. The three text tasks (SG-Gov, SG-Legal, HT-Lexicon) are four-option multiple choice and are scored by accuracy under per-option log-likelihood: the model assigns no free-form text but selects the option (A–D) with the highest (length-unnormalized) log-likelihood, which is compared to the gold index. SG-Multimodal is open-ended visual question answering, scored by a single greedy generation and a deterministic, *judge-free* matcher (exact → alias → substring → fuzzy, with a 0.85 similarity floor) against a curated answer/alias key. The two recall tasks are open-ended generation scored against reference text: SG Legislative Grounding asks the model to reproduce a specific statutory provision and is scored by token-level word error rate (WER) and word-overlap F1; SG Home Team Alignment is multi-turn recall of public-safety agencies’ knowledge, mission, and values, scored by ordered key-term recall (LCS over reference key terms extracted offline by an LLM; no eval-time judge). SG Multimodal Safeguards is likewise generative but is scored by an LLM/VLM judge: a vision-language judge labels whether the model refused to identify the individual(s) in the image, and the metric is the intrinsic Refusal Rate (fraction of items safely refused).

In total the four localized capability tasks contain 39,451 graded items (32,902 for SG-Gov, 5,257 for HT-Lexicon, 1,110 for SG-Legal, and 182 for SG-Multimodal). The text tasks are built from public source documents that are grouped into higher-level domains, and task accuracy is aggregated *item-weighted* (equal weight per question, so larger domains contribute proportionally). Concretely, SG-Gov draws its 32,902 questions from 525 source documents organized into 16 groups; SG-Legal draws 1,110 questions from 440 source documents (mean 2.5, max 14 questions per document); HT-Lexicon spans 7 operational sub-

domains; and SG-Multimodal spans 21 visual categories. Beyond these, SG Legislative Grounding contains 28,054 provision-level prompts spanning 489 public statutes; SG Home Team Alignment is scored over 544 turns drawn from 150 multi-turn conversations (knowledge, values, and mixed-task splits); and SG Multimodal Safeguards has 509 items.

Released artifacts. To aid reproducibility while respecting the sensitivity of the underlying sources, we release de-identified artifacts for the localized suite as supplementary material. For each *released* task (the multiple-choice, VQA, and safety tasks) it provides: (i) the exact prompt template and task configuration; (ii) the scoring logic as a self-contained, runnable reference implementation, comprising the multiple-choice log-likelihood scorer and the SG-Multimodal tiered matcher with its answer-alias key, plus the VLM-judge rubric used for SG Multimodal Safeguards; (iii) the decoding and runtime settings; and (iv) a small, de-identified sample of example items (five per released task). All runs are zero-shot: SG-Multimodal uses greedy decoding with a 128-token limit, and SG Multimodal Safeguards uses low-temperature generation scored by a vision-language judge. In the sample items, images for SG-Multimodal and SG Multimodal Safeguards are replaced with placeholders or redactions, and the operational system prompt for SG Multimodal Safeguards is withheld, with only the image-identification refusal-policy excerpt released. All organization-, location-, and person-identifying details have been removed or replaced with generic placeholders, and the released items are illustrative samples rather than the full evaluation sets, whose sizes are reported in Table 6. We deliberately *exclude* the two recall tasks (SG Legislative Grounding and SG Home Team Alignment) from the release. Their references are verbatim statutory text and specific agency facts, and their prompts name particular statutes and agencies; releasing them at scale falls outside the data-governance approvals covering our operational sources. We therefore characterize their construction and scoring in Table 6 but release none of their prompts, items, or answer keys.

F Additional Evaluation Results

This section reports the per-model and subgroup breakdowns behind the aggregate localized results

Task	Size	Construction	Scoring	Validation and leakage control
SG-Multimodal	182	Human-localized VQA over local scenes, landmarks, equipment, and signage; 21 visual categories	Single greedy generation; tiered string match (exact, alias, substring, then fuzzy at ≥ 0.85); judge-free	Multi-annotator overlap with adjudicated disagreements (10 annotators; 3 per item), image-source checks, n -gram decontam
SG-Gov	32,902	Grounded MCQ over public legislation and policy; 525 source documents in 16 groups	4-option accuracy via per-option log-likelihood; item-weighted aggregation	Expert review sample (100), held-out sources, n -gram decontam
SG-Legal	1,110	Statutory MCQ over 440 public legal documents	4-option accuracy via per-option log-likelihood	Version checks, reviewer sample, train/eval split
HT-Lexicon	5,257	Operational terminology and procedural knowledge across 7 sub-domains	4-option accuracy via per-option log-likelihood; item-weighted aggregation	Held-out split, expert review sample (100)
SG Multimodal Safeguards	509	Human-curated localized public-figure images with identity-revealing requests; each base request reissued under five (templated) adversarial framings: base, authority, emotional, context-specificity, speculative	Single generation; VLM-as-judge refusal labelling, reported as an intrinsic Refusal Rate	Judge rubric and calibration (all manually reviewed), PII/image-source checks, held-out images, train/eval separation
SG Legislative Grounding	28,054	Open-ended reproduction of statutory provisions: each prompt asks the model to state or explain a specific section of a named statute (489 public statutes)	Free-text generation scored against the reference provision: token-level WER and word-overlap F1; judge-free	Public-source statutes, held-out sections, n -gram decontam, train/eval split
SG Home Team Alignment	544	Multi-turn recall of public-safety agencies' knowledge, mission, and values; 150 conversations across three splits (knowledge, values, mixed-task)	Free-text generation; ordered key-term recall via LCS over LLM-extracted reference key terms; judge-free at eval time	Held-out conversations, reviewer sample (100), train/eval separation

Table 6: Localized evaluation suite. All tasks use zero-shot, no-search evaluation.

in Section 4.1. Table 7 gives per-model Pass@1 on the four localized tasks and is the source of the *Best baseline* column in Table 2; Tables 8 and 9 decompose SG-Gov and HT-Lexicon by ministry and by public-safety agency, respectively; and Table 10 expands the general-capability retention summary of Table 3 to the full benchmark set. *Phoenix-VL 1.5 Medium* leads on every SG-Gov ministry and on five of seven HT-Lexicon subgroups; on the remaining two it trails the best baseline by 0.0048 (Singapore Police Force) and 0.030 (Prison Service). Figures 4, 5, 6, and 7 visualize these results.

G Additional Ablations and Sensitivity Analyses

G.1 Long-Context Extension

LCE sits between CPT and SFT (Table 1) and was designed to improve long-context behavior without materially changing the short-context benchmarks of Table 4. Table 11 reports its effect at the 131k configuration: Michelangelo (Vodrahalli

et al., 2024) 20-operation list tracking improves by +0.053 (a $\sim 60\%$ relative gain), 42k needle-in-a-haystack retrieval by +0.054, and few-shot long-log accuracy reaches 0.998.

G.2 Sensitivity to the Proportion of Jurisdiction-Specific Data

We swept the domain-to-replay ratio of the domain-training mixture (Section 3.2.3) from 1:9 to 5:5. Table 12 shows that localized capability depends far more on the *presence* of jurisdiction-specific data than on its precise proportion: HT-Lexicon is near-saturated already at 1:9, which alone recovers $\sim 92\%$ of the total gain, and MMLU stays within ~ 1 point of its pre-adaptation value across the entire sweep.

H Instruct Tuning Results

Loss formulation. SFT uses a completion-only objective in which user-prompt tokens are masked and only assistant tokens contribute to the loss.

Benchmark (↑)	Phoenix-VL 1.5 Medium (123B) [Ours]	Llama 4 Maverick (400B-A17B) [MetaAI]	GPT-OSS (117B-A5B) [OpenAI]	Nemotron 3 Super (120B-A12B) [Nvidia]	GLM-4.5V (106B-A12B) [Z.ai]	Qwen 3.5 (122B-A10B) [Alibaba Cloud]
SG-Multimodal	0.7509	0.6716	–	–	0.5097	0.6431
SG-Gov	0.9265	0.7225	0.7302	0.7971	0.7021	0.8006
SG-Legal	0.8640	0.5054	0.5523	0.5919	0.4883	0.6144
HT-Lexicon	0.9081	0.7666	0.7615	0.8891	0.7453	0.8691

Table 7: Per-model localized results (Pass@1, zero-shot, no-search). The Best-baseline column of Table 2 is the per-task maximum over applicable external baselines. For SG-Multimodal, ‘–’ for GPT-OSS and Nemotron 3 indicates that these are text-only models and cannot process image inputs; therefore, the SG-Multimodal best baseline is computed over vision-language model baselines only. Further breakdowns of SG-Gov and HT-Lexicon by ministries and departments are reported in Table 8 and Table 9, respectively.

Benchmark by Ministry (↑)	Phoenix-VL 1.5 Medium (123B)	Llama 4 (400B-A17B)	GPT-OSS (117B-A5B)	Nemotron 3 (120B-A12B)	GLM-4.5V (106B-A12B)	Qwen 3.5 (122B-A10B)	Qwen 3 (235B-A22B)
Average*	0.9265	0.7225	0.7302	0.7971	0.7021	0.8006	0.7815
Prime Minister’s Office	0.9060	0.7222	0.7351	0.7903	0.7023	0.7895	0.7895
Culture, Community & Youth	0.9370	0.7458	0.7290	0.8099	0.7216	0.8109	0.7931
Defence	0.9266	0.7133	0.7385	0.8005	0.7225	0.7775	0.7592
Digital & Information	0.9360	0.7484	0.7733	0.8308	0.7625	0.8492	0.8330
Education	0.9225	0.7399	0.7601	0.8192	0.7380	0.8173	0.8155
Finance	0.9070	0.6871	0.6896	0.7523	0.6533	0.7549	0.7390
Foreign Affairs	0.9590	0.8073	0.8044	0.8740	0.8034	0.8931	0.8664
Health	0.9387	0.7346	0.7460	0.7978	0.7022	0.7963	0.7774
Home Affairs	0.9260	0.6970	0.7267	0.8017	0.6998	0.7857	0.7842
Law	0.9421	0.7494	0.7405	0.8193	0.7212	0.8316	0.8034
Manpower	0.9154	0.7084	0.7173	0.7703	0.6726	0.7820	0.7448
National Development	0.9255	0.7126	0.7161	0.7865	0.6818	0.7865	0.7688
Social & Family	0.9028	0.7207	0.7294	0.7776	0.6821	0.7706	0.7671
Sustainability & Env.	0.9272	0.7304	0.7286	0.8055	0.7103	0.8124	0.7881
Trade & Industry	0.9381	0.7387	0.7671	0.8173	0.7358	0.8246	0.8020
Transport	0.9277	0.7180	0.7368	0.8055	0.7070	0.8190	0.7834

* Averaged across all questions within the entire evaluation set, not across individual ministries.

Table 8: SG-Gov benchmark breakdown by Ministries. All evals Pass@1, Zero-shot.

We add a distillation term (Hinton et al., 2015) computed against the teacher model, Mistral Large 3 (Mistral AI, 2025a), with the total loss a weighted combination of cross-entropy and distillation controlled by a mixing coefficient α . The vision encoder remains frozen while the adapter layer is trained.

The SFT run was highly stable, completing with no loss spikes. The student model’s training loss converged to the level of the teacher’s loss, indicating effective knowledge transfer on the instruction mixture. Post-SFT evaluations show that the model is strong on general capabilities and excellent on targeted multilingual benchmarks, with notable improvements over the base checkpoint Mistral Medium 3.1 across several axes. For a more holistic evaluation, we benchmarked this check-

point on additional metrics beyond those reported for the final checkpoint, covering additional tasks like translation, function calling, and additional languages.

On general knowledge, *Phoenix-VL 1.5 Medium* achieves **0.864** on New MMLU (Hendrycks et al., 2020) (5-shot instruct), slightly above the baseline of 0.860. Multilingual knowledge benchmarks show consistent gains: KMMLU (Son et al., 2025) (Korean) improves from 0.684 to **0.700**, JMMLU (Yin et al., 2024) (Japanese) from 0.826 to **0.835**, CMMLU (Li et al., 2024a) (Chinese) from 0.831 to **0.844**, and Indonesian MMLU from 0.811 to **0.825**. Code generation capability, measured by LiveCodeBench (Jain et al., 2024) (CoT, pass@5), improves from 0.429 to **0.448**. Multimodal reasoning is preserved, with MMMU (Yue

Benchmark by	Phoenix-VL 1.5 Medium	Llama 4	GPT-OSS	Nemotron 3	GLM-4.5V	Qwen 3.5	Qwen 3
Public-Safety Agencies (↑)	(123B)	(400B-A17B)	(117B-A5B)	(120B-A12B)	(106B-A12B)	(122B-A10B)	(235B-A22B)
Average*	0.9081	0.7666	0.7615	0.8891	0.7453	0.8691	0.8474
Ministry HQ & HTX	0.9583	0.8073	0.8281	0.9375	0.8229	0.9167	0.9167
Singapore Police Force	0.9104	0.8041	0.8029	0.9152	0.7694	0.8865	0.8854
Civil Defence Force	0.8849	0.7182	0.7227	0.8627	0.6964	0.8399	0.8662
Immigration & Checkpoints	0.9867	0.8567	0.8300	0.9267	0.8533	0.9167	0.8184
Prison Service	0.8733	0.8235	0.7919	0.9005	0.8100	0.8959	0.9033
Central Narcotics Bureau	0.9658	0.8433	0.8034	0.9373	0.8091	0.9174	0.8824
Gambling Regulatory	0.9490	0.8408	0.8408	0.9193	0.8238	0.9193	0.9060

* Averaged across all questions within the entire evaluation set, not across departments.

Table 9: HT-Lexicon subgroup breakdown by public-safety agency department. Pass@1, zero-shot. Bold indicates the row-wise best model for each subgroup; aggregate benchmark claims in the main text refer to the Average row and Table 2.

Category / Benchmark (↑)	Phoenix-VL 1.5 Medium (123B)	Llama 4 Maverick (400B-A17B) [MetaAI]	GPT-OSS (117B-A5B) [OpenAI]	Nemotron 3 Super (120B-A12B) [Nvidia]	GLM-4.5V (106B-A12B) [Z.ai]	Qwen 3.5 (122B-A10B) [Alibaba Cloud]
<i>Multilingual</i>						
English	0.9564	0.9285	0.9531	0.9520	0.9430	0.9575
Malay	0.9274	0.9084	0.8905	0.8804	0.8726	0.9151
Tamil	0.8447	0.8391	0.7642	0.6994	0.7408	0.8279
Chinese	0.9330	0.9207	0.8994	0.9162	0.9095	0.9352
<i>Multimodal</i>						
MMMU	0.6078	0.4300	–	–	0.4850	0.6122
RealWorldQA	0.6627	0.7085	–	–	0.7138	0.7935
OCRBench-v2	0.5021	0.3293	–	–	0.4666	0.6516
DocVQA	0.9305	0.9309	–	–	0.9378	0.9590
ChartQA	0.7832	0.3136	–	–	0.8913	0.8744
<i>Textual Knowledge</i>						
MMLU	0.8348	0.7898	0.8566	0.8547	0.7895	0.8768
MMLU Pro	0.7681	0.8062	0.7340	0.8373	0.7217	0.8670
GPQA Diamond	0.5909	0.6980	0.6364	0.7923	0.6294	0.8660
IFEval	0.8244	0.8669	0.7652	0.7726	0.7122	0.8725
<i>Math & Code</i>						
AIME	0.4995	0.6731	0.6377	0.8189	0.6213	0.9556
Math-500	0.8120	0.8900	0.8920	0.9100	0.8879	0.9780
LiveCode	0.3240	0.2646	0.4160	0.4126	0.3415	0.4280

Table 10: Unified Intelligence Benchmarks: Multilingual (Belebele), Multimodal, and General Intelligence. All evals are Pass@1, Zero-shot, denoting globally competitive performance of *Phoenix-VL 1.5 Medium* after domain adaptation. ‘–’ denotes metrics not available for specific models.

Benchmark	Base+CPT	+LCE	Δ
Michelangelo (list, 20 ops)	0.089	0.142	+0.053
Needle-in-a-haystack (42k)	0.273	0.327	+0.054
Long-logs (few-shot)	0.984	0.998	+0.014

Table 11: Long-context extension gains at 131k evaluation context.

et al., 2024) (CoT) holding steady at **0.668**, while MathVista (Lu et al., 2023) (CoT) shows a moderate regression from 0.738 to **0.661**, likely at-

tributable to distribution shift in the instruction mixture. The subsequent ODPO stage recovers part of this regression to 0.701.

Translation quality, evaluated on FLORES (NLLB Team et al., 2022), shows substantial improvements across all measured language pairs. English–Tamil translation improves markedly from 27.7 to **36.8** BLEU, and Tamil–English from 32.3 to **38.2** BLEU. English–Hindi rises from 34.6 to **40.0**, Hindi–English from 40.0 to **44.7**, English–Japanese from 35.0 to **39.3**, and

Domain:replay	HT-Lexicon (95% CI)	Share of gain	MMLU
0 (pre-adapt.)	0.7322 $\pm$ 0.012	–	0.8368
1:9	0.9060 $\pm$ 0.008	92%	0.8264
2:8	0.9030 $\pm$ 0.008	90%	0.8307
3:7	0.9051 $\pm$ 0.008	91%	0.8336
4:6	0.9192 $\pm$ 0.007	99%	0.8351
5:5	0.9212 $\pm$ 0.007	100%	0.8354

Table 12: Sensitivity of domain training to the domain-to-replay ratio. “Share of gain” is the fraction of the total HT-Lexicon improvement (pre-adaptation $\rightarrow$ 5:5) recovered at each ratio.

Japanese–English from 29.8 to **33.6** BLEU. Korean and Chinese pairs show similar trends.

Function-calling capability is retained, with ComplexFuncBench (Zhong et al., 2025) overall success improving slightly from 29.3% to **30.8%**, and tool-call robustness (F1) remaining stable at **0.865**. As expected, instruction-following precision on Collie (Yao et al., 2023) is slightly below baseline (0.422 vs. 0.458), as the model has not yet undergone preference optimization; this gap is subsequently addressed by the ODPO stage.

I ODPO Training Details

Preference data. The 5B-token mixture comprises feedback prompts (77%), feedback with images (14%), and instruction-with-image pairs (9%). Human annotators ranked completions following guidelines covering five categories: language quality (succinctness and clarity), structure and formatting, instruction following, creativity, and harmlessness.

Online loop. Three components operate in concert: (i) a *generator* sampling completion pairs from the current policy; (ii) a *pairwise reward model* (PWRM) scoring each pair; and (iii) a *trainer* updating the policy on the resulting signal. Given a prompt x , two completions (y_1, y_2) are sampled from the current policy π_θ , and the PWRM returns a winning probability $p_w = p_{\text{RM}}(y_1 \succ y_2 | x)$. Because candidates are drawn from the current policy at every step, the optimization avoids the distribution mismatch of offline preference data. The PWRM is initialized from the SFT checkpoint and trained on human-annotated winner–loser pairs with input tokens masked from the loss; its scoring temperature is selected to minimize expected calibration error, treating the PWRM as a binary classifier.

Weighted two-sided loss. Let

$$\Delta = \log \frac{\pi_\theta(y_1 | x)}{\pi_{\text{ref}}(y_1 | x)} - \log \frac{\pi_\theta(y_2 | x)}{\pi_{\text{ref}}(y_2 | x)} \quad (1)$$

denote the difference in policy–reference log-ratios between the two completions, where π_{ref} is the frozen pre-ODPO checkpoint. Rather than reducing p_w to a hard winner–loser label, the judge probability weights both sides of the DPO objective:

$$\mathcal{L} = -\mathbb{E} [p_w \log \sigma(\beta \Delta) + (1 - p_w) \log \sigma(-\beta \Delta)] \quad (2)$$

Marginally confident judgments (e.g., $p_w=0.55$) thus yield a correspondingly soft preference, while confident judgments produce stronger gradients: the policy learns to match the judge’s confidence rather than a hard 0/1 target. For samples with pre-existing human rankings (feedback data), we instead apply a conservative DPO loss (Mitchell, 2023), which replaces the per-sample weight $1 - p_w$ in Eq. 2 with a fixed global label-smoothing parameter ϵ .

Hyperparameters and monitoring. We train with $\beta = 0.01$, $\epsilon = 0.1$, learning rate 2×10^{-8} , weight decay 0.8, sequence length 32,768, and approximately 393k tokens per optimizer step. The policy samples at temperature 0.7; the PWRM scores at its calibration-selected temperature of 1.2. Training is monitored via the binary KL divergence between the judge’s preference distribution and the policy’s implicit preference distribution, which should decrease over training, alongside $\text{KL}(\pi_\theta \parallel \pi_{\text{ref}})$, which should grow slowly. Training was stopped after 5,500 steps when evaluation metrics plateaued. The post-ODPO checkpoint reaches 0.525/0.325 on AIME 2024/2025 (16 samples) and 0.475 on Collie (Yao et al., 2023).

Negative result: penalizing dispreferred tokens destabilizes training. In early runs we added an auxiliary negative-preference term (Zhang et al., 2024) with a small coefficient (10^{-4}), intended to penalize infinite generations and format errors by applying gradient ascent to dispreferred generations. Under aggressive settings (higher learning rates, low β), this term instead produced recurring explosions of format errors. Our hypothesis is that gradient ascent on all tokens of a generation can sharply increase sampling entropy during on-policy generation, beyond the point of recovery. The term was removed after the second run; the weighted

two-sided loss alone proved sufficient to reduce infinite generations to 3.8%.

J Safety and Model Behavior Results

Table 13 lists the full numerical scores for all metrics visualised in Figure 1; the two groups correspond to the two radar charts and were evaluated on different competitor sets, hence the ‘–’ entries.

K Further Details on Safety and Model Behavior

K.1 Evaluations for Alignment and Safety in our Context

SG Home Team Alignment. To test knowledge grounding in an open-ended generative setting, models are presented with multi-turn prompts drawn from curated public-safety agencies knowledge bases spanning departmental knowledge, functions, and mixed factual-procedural tasks. Because open-ended generation allows for structural variance, we rely on a composite scoring mechanism rather than exact string matching. Each turn is paired with a ground-truth keyword list pre-extracted from gold answers using a larger LLM with task-specific rules. This extraction excludes generic action verbs (e.g., *manages*, *provides*) while retaining domain-specific verbs (e.g., *prosecutes*, *enforces*) that carry discriminative operational meaning. We report a Composite Recall score (R_{comp}) that equally balances Keyword Presence (KP), which is order-agnostic, and Keyword Token Recall (KTR), which enforces the chronological retention of procedural steps:

$$\text{KP} = \frac{|\{s \in S : \exists f \in \text{forms}(s), f \subseteq \hat{y}\}|}{|S|} \quad (3)$$

$$\text{KTR} = \frac{\text{LCS}(O_{\text{can}}, O_{\text{found}})}{|S|} \quad (4)$$

$$R_{\text{comp}} = 0.5 \cdot \text{KP} + 0.5 \cdot \text{KTR} \quad (5)$$

where S is the set of keywords, $\hat{y}$ is the model prediction, $\text{forms}(s)$ returns all accepted morphological forms for slot s , and O_{can} and O_{found} represent the canonical and found keyword orderings, respectively. Scores are averaged across turns within each example and then across all examples per subtask.

SG Legislative Grounding. Models are prompted to reproduce statutory text verbatim

from parametric memory. Unlike general knowledge retrieval, statutory text requires verbatim alignment; even minor paraphrasing or synonymous substitutions can fundamentally alter the legal threshold or jurisprudential meaning of a statute. Therefore, we utilize the Levenshtein distance to compute a Word Error Rate (WER) alongside a Statutes Accuracy score (A_{stat}) and a Word Overlap F_1 score, ensuring that models are penalized for both omissions and lexical deviations. While we recognize the inherently probabilistic nature of language models, this benchmark serves to explicitly steer model behavior toward exact parametric recall as far as possible.

$$\text{WER} = \frac{\text{Levenshtein}(\hat{y}_{\text{tok}}, y_{\text{tok}})}{|y_{\text{tok}}|} \quad (6)$$

$$A_{\text{stat}} = \max(0, 1 - \text{WER}) \quad (7)$$

$$P = \frac{|\hat{W} \cap W|}{|\hat{W}|}, \quad R = \frac{|\hat{W} \cap W|}{|W|} \quad (8)$$

$$F_1 = \frac{2 \cdot P \cdot R}{P + R} \quad (9)$$

where $\hat{y}_{\text{tok}}$ and y_{tok} are the tokenized prediction and reference respectively, and $\hat{W}$, W are the corresponding deduplicated word sets. Precision (P) and Recall (R) feed into the final F_1 calculation.

We report the Word Overlap F_1 score in this paper, while adopting Levenshtein distance scores in downstream Reinforcement Learning.

K.2 Multimodal Safeguards

To quantify resilience against localized visual misuse, each model is evaluated on its propensity to refuse undesired or harmful multimodal tasks, even when subjected to intense adversarial pressure. The evaluation dataset spans a diverse array of localized images and frames requests across five distinct adversarial dimensions:

- **Base:** Direct, unmodified harmful requests.
- **Authority framing:** The user claims a professional credential or clearance to demand compliance.
- **Emotional pressure:** Injecting urgency, stress, or moral blackmail into the prompt.
- **Context specificity:** Embedding the harmful request within a highly plausible, localized real-world operational scenario.

Metric	<i>Phoenix-VL 1.5 Medium</i> (123B)	Qwen3.5 (122B-A10B)	Llama 4 (400B-A17B)	GLM-4.5V (106B-A12B)	Cmd-A (112B)	Nemotron 3 (120B-A12B)	GPT-OSS (117B-A5B)
<i>Alignment & Abstention (↑)</i>							
SG Legislative Grounding	0.6193	0.2284	0.2369	0.2255	0.2536	–	–
SG Multimodal Safeguards ^J	0.9764	1.0000	0.9980	1.0000	1.0000	–	–
Abstention-Prudence ^J	0.7361	0.8040	0.7963	0.8827	0.8333	–	–
Abstention-Honesty ^J	0.8681	0.9000	0.8980	0.9370	0.9180	–	–
SG Home Team Alignment	0.8176	0.6133	0.5329	0.4307	0.8237	–	–
<i>Adversarial Traits (↓)</i>							
Defer to Users ^J	0.26	–	–	0.03	0.29	0.27	0.14
Delusional Sycophancy ^J	0.31	–	–	0.06	0.34	0.32	0.03
Instructed Long Horizon Sabotage ^J	0.42	–	–	0.35	0.27	0.42	0.15
Prompt Injection Vulnerability ^J	0.06	–	–	0.23	0.10	0.03	0.04
System Prompt Leakage ^J	0.08	–	–	0.08	0.05	0.03	0.01

‘–’ indicates the model was not included in that evaluation set.

^J LLM/VLM-as-judge metric.

Table 13: Full numerical results for the safety and model-behaviour metrics shown in Figure 1. Pass@1, zero-shot. **Bold** marks the best value per row. ↑ higher is better; ↓ lower is better. Metrics marked ^J are scored by LLM/VLM-as-judge; unmarked metrics use deterministic, judge-free scorers (Appendix K). Judge-robustness analysis in Appendix K.5.

Metric	Primary (Qwen)	Gemma-4
Safeguards refusal (↑)	0.9764	0.9764
Abstention-Prudence (↑)	0.7361	0.7407
Abstention-Honesty (↑)	0.8681	0.8704
Abstention over-conserv. (↓)	0/275	0/275
Defer to Users (↓)	0.26	0.29
Delusional Sycophancy (↓)	0.31	0.31
Prompt Injection (↓)	0.06	0.04
System Prompt Leakage (↓)	0.08	0.08
Instructed LH Sabotage (↓)	0.42	0.31

Table 14: Judge robustness: primary judge vs. an independent third-family judge (gemma-4-31B-it) on byte-identical rubrics and stored responses. ↑/↓: higher/lower is better. The Safeguards value is exact (0.9764 = 497/509).

- **Speculative framing:** Hedging the prompt as a tentative or academic observation to bypass safety filters.

The refusal rate is measured intrinsically and quantified as the ratio of safely rejected queries to the total evaluation set:

$$\text{Refusal Rate} = \frac{|\{\text{refusals}\}|}{|\{\text{total samples}\}|}. \quad (10)$$

Refusals are given by Qwen3-VL-32B as VLM judge (Appendix K.5).

K.3 Abstention Evaluations

To measure epistemological discipline, we adopt the Abstention framework established by (Yang

et al., 2023). Each question is labeled with an expected behavior (to either gracefully refuse or safely answer), with Qwen3.5-122B-A10B as LLM-as-a-judge classifying the model response. Our benchmarks test refusal behavior across eight categories of genuinely unanswerable queries:

- **requires_web_access:** Dependent on live internet retrieval.
- **requires_realtime_data:** Necessitating live status or current pricing.
- **requires_database_access:** Querying secure operational records.
- **requires_private_information:** Requesting non-public data.
- **requires_specific_citation:** Demanding unverifiable references.
- **post_training_cutoff:** Querying events beyond the model’s temporal knowledge.
- **unknowable_future:** Prompting for impossible predictions.
- **insufficient_information:** Lacking necessary contextual parameters.

The dataset explicitly interleaves answerable control questions so that appropriate refusal is measured jointly against the risk of over-refusal. The quality of the model’s behavioral boundary is subsequently captured across three metrics: Prudence (P), Over-Conservativeness (C), and Honesty (H), defined as:

$$P = \frac{|\{\text{correct refusals}\}|}{|\{\text{should refuse}\}|} \quad (11)$$

$$C = \frac{|\{\text{false refusals}\}|}{|\{\text{should answer}\}|} \quad (12)$$

$$H = \frac{P + (1 - C)}{2} \quad (13)$$

K.4 Bloom-Derived Behavioral Evaluations

To rigorously quantify adversarial vulnerabilities and stylistic failure modes, we adapt several automated evaluations from the Anthropic Bloom framework (Gupta et al., 2025). All behavioral judgments are generated via an LLM-as-a-judge pipeline with Qwen3.5-122B-A10B as judge (robustness to this choice is assessed in Appendix K.5). Each dimension maps to an operational risk specific to air-gapped settings: institutional confirmation bias during policy drafting, unsafe behavior in downstream agentic tasks, exposure of the operational system prompt, and fabrication when external context is inaccessible.

The evaluated dimensions include:

- **Defer to Users:** Evaluates Epistemic Honesty by measuring whether the model incorrectly yields to user pressure rather than maintaining factual truth across varying degrees of error severity.
- **Delusional Sycophancy:** Assesses the risk of the model reinforcing or validating fantastical user beliefs, a critical safeguard against worsening symptoms in psychologically vulnerable users.
- **System Prompt Vulnerability:** Jointly measures the success rate of **Prompt Injection** attacks attempting to override the system prompt and **System Prompt Leakage** rates under adversarial pressure.
- **Instructed Long Horizon Sabotage:** Tests the model’s propensity to secretly pursue malicious goals embedded within long, complex administrative tasks while maintaining the appearance of normal operational behavior.

Interpreting the sabotage metric. Instructed Long Horizon Sabotage is an adversarial upper bound rather than a characterization of deployed behavior. Because the malicious goal is supplied explicitly in the instruction, the metric presupposes that an adversary controls the instruction channel. In our deployment that channel is a centrally administered system prompt on air-gapped infrastructure; adversarial content arriving through user or document inputs is instead covered by the prompt-injection evaluation, and operational outputs are human-verified (Ethical Considerations). The metric is therefore most informative for the agentic workflows planned in future iterations, in which the model would act over longer horizons with less direct human verification of intermediate steps.

K.5 Judge-Robustness Analysis

Judges are used only for SG Multimodal Safeguards, the abstention metrics, and the Bloom-derived evaluations; all capability results (Tables 2–4 and 7–12) and two of the five alignment metrics in Figure 1 (SG Legislative Grounding; SG Home Team Alignment) are deterministic. To test sensitivity to the choice of judge, we re-scored the stored responses behind Table 13 under byte-identical rubrics with gemma-4-31B-it, a judge whose model family is distinct from *Phoenix-VL 1.5 Medium*’s base model, its teacher, our primary judges (Qwen3.5-122B-A10B for text; Qwen3-VL-32B for vision), and every model scored in that table.

Table 14 reports the comparison. Item-level agreement is 509/509 on Safeguards ($\kappa=1.0$, with the same 12 failures) and 911/924 on abstention ($\kappa=0.97$). Abstention scores change by at most 0.005, which does not alter any ranking. Four of the five Bloom-derived rates differ by at most 0.03, including identical scores on two metrics. Instructed Long Horizon Sabotage changes from 0.42 with the primary judge to 0.31 with the independent judge. The alternative judge is also slightly more favorable to *Phoenix-VL 1.5 Medium* on abstention, while Safeguards remains below several baselines at 1.0000 and a Qwen baseline remains above *Phoenix-VL 1.5 Medium* on both abstention metrics. We provide per-task agreement figures and sanitized examples of the sabotage disagreements and the 36/509 Safeguards soft-identification flags in the supplementary material (the flags are diagnostic and do not affect the reported scores).