AI Model Selection Quick Reference

For Aid4Mail practitioners and team leads

Use this guide to choose a practical AI model for Aid4Mail filtering, classification, and review workflows. For an in-depth user guide, read our Aid4Mail AI Email Review Workflow Guide.

1. Fastest Practical Recommendations

Main priority	Start with	Why	Main caution
Best practical offline default for high-volume binary review	**Mistral Small 3.2 24B**	Best balance of speed, accuracy, privacy, cost control, deterministic behavior, and long-term model control on consumer 24–32 GB GPUs. Processes about **614,000 emails per 62-hour weekend** in Test 5 conditions.	Validate language support and monitor the INCONCLUSIVE queue.
Too many Mistral INCONCLUSIVE results	**Second pass with Qwen 3.6 27B Dense**	Use the slower, more consistent multilingual model only on the gray-zone subset, not the full corpus.	Slow: not suitable as the first pass for large corpora on the reference workstation.
Faster second-pass alternative	**Qwen 3.6 35B MoE**	Similar family strengths to Qwen Dense, roughly twice as fast in Test 1, with a small accuracy concession.	Still much slower than Mistral Small.
Lowest-cost cloud classification	**Gemini 3.1 Flash-Lite**	Lowest retained cloud cost, strong multilingual and multi-category results.	Treat as non-deterministic; verify region and provider terms before use.
Highest cloud accuracy / decision coverage	**Grok 4.2 Non-Reasoning**	Top cloud model for Test 1 F1 and tied for top Automation Yield. Deterministic retained cloud option.	Higher cloud cost than Gemini Flash-Lite; cloud data-transfer issues still apply.

Premium analysis, translation, or reasoning-heavy work	**Claude 4.7 Opus**	Strong for analysis beyond classification, including nuanced language and reasoning tasks.	Slowest and most expensive retained cloud model; not the routine classification default.
Accuracy-first multilingual offline work	**Qwen 3.6 27B Dense**	Only offline model with 100% on Tests 2, 3, and 4; narrowest spread across the four accuracy tests.	Very low throughput on the reference workstation.
Fastest raw offline triage	**Ministral 3 14B**	Fastest retained model, about **879,000 emails per weekend** in Test 5 conditions, and runs on lower-end 16 GB GPUs.	Weaker for broad multi-category discrimination.
High English offline decision coverage on high-end hardware	**Llama 3.3 70B**	Strong English accuracy and tied top Automation Yield.	Needs about 80 GB VRAM for efficient full-speed operation; slow on 32 GB with CPU offload.

2. Know the Three Numbers That Matter

Metric	Plain-language meaning	Why it matters
F1	How good the model's committed decisions are.	Useful when you can review ambiguous items separately.
Automation Yield (AY)	How much of the full corpus the model resolved correctly without deferring.	Useful when every INCONCLUSIVE item creates review work.
INCONCLUSIVE rate	How often the model says the email is too ambiguous to decide safely.	A small INCONCLUSIVE queue can be a safety feature; a large one can become an operational burden.

Do not choose by a single score. **Mistral Small 3.2 24B leads Test 1 F1 on decided emails**, but it also produced more INCONCLUSIVE results than other retained models. That does not make it weak; it means it uses the three-label boundary more conservatively. In many forensic and eDiscovery workflows, that is preferable to forcing borderline emails into Responsive or Unresponsive.

3. First Decision: Cloud, Enterprise Cloud, or Offline?

If email data cannot leave your environment

Use an **offline model** through Ollama or LM Studio.

Recommended starting points:

- **Mistral Small 3.2 24B** for high-volume binary responsiveness, insider-threat, exfiltration, and similar focused classification work.
- **Qwen 3.6 27B Dense** for accuracy-first multilingual or multi-category work on small-to-medium corpora.
- **Qwen 3.6 35B MoE** when Qwen Dense is too slow but Qwen-family multilingual strength is needed.
- **Gemma 4 26B** for balanced offline multilingual work, especially where Korean support matters.
- **Ministral 3 14B** when 16 GB VRAM or maximum raw speed is the constraint.
- **Llama 3.3 70B** only when high-end VRAM is available or slower overnight/weekend throughput is acceptable.

Offline advantages:

- Email content stays local.
- No cloud token charges.
- No cloud rate limits.
- No provider retirement risk for the downloaded model version.
- Better long-term control over a repeatable workflow.

Offline does still require local security, hardware sizing, validation, and preservation of outputs.

If cloud is allowed but governance or region matters

Use an **enterprise cloud platform** rather than a simple direct API when you need regional deployment, explicit quotas, enterprise billing, stronger governance, or compliance documentation.

Practical regional notes:

- **Gemini 3.1 Flash-Lite** on the `eu` multi-region endpoint excludes the **UK and Switzerland**.
- For Swiss or UK residency on a premium cloud model, the cited option is **Claude 4.7 Opus via Amazon Bedrock**.
- When regional requirements are strict and no approved cloud region fits, use offline processing.

Always verify current provider availability, data-use terms, logging, retention, and contractual protections before processing live matter data.

If cloud is allowed without special restrictions

Choose by task:

- **Lowest cost:** Gemini 3.1 Flash-Lite.
- **Best cloud accuracy / coverage:** Grok 4.2 Non-Reasoning.
- **Premium analysis and translation:** Claude 4.7 Opus.
- **Broad multilingual or multi-category cloud triage:** Gemini 3.1 Flash-Lite.

---

4. Recommended Choices by Priority

Cost

Best low-cost cloud: Gemini 3.1 Flash-Lite.

It is the cheapest retained cloud option at roughly **$42–$43 per 100,000 emails** in Test 5 conditions. Use the Agent Platform deployment when the matter is eligible for the US or EU multi-region endpoint and speed matters.

Best zero-marginal-token-cost option: Offline models.

Offline is not automatically cheaper for every low-volume matter because hardware, electricity, and IT support still exist. It becomes especially attractive when:

- data must stay local;
- workloads are recurring or high-volume;
- rate limits are unacceptable;
- long-term access to the same model version matters;
- avoiding per-token approval and billing is operationally useful.

Speed

Use two categories, not one:

- **Maximum raw speed:** Ministral 3 14B.
- **Best speed with top-tier binary accuracy:** Mistral Small 3.2 24B.

For production classification where accuracy matters, **Mistral Small 3.2 24B is the stronger practical default**. Ministral is useful for fast coarse triage, constrained hardware, or lower-risk passes, but it is weaker for broad multi-category discrimination.

Accuracy and decision coverage

Use the model that matches the accuracy problem:

- **Highest Test 1 F1 on decided emails:** Mistral Small 3.2 24B.
- **Highest cloud F1 and top Automation Yield:** Grok 4.2 Non-Reasoning.
- **Highest offline multilingual consistency:** Qwen 3.6 27B Dense.
- **High English offline decision coverage:** Llama 3.3 70B, if suitable hardware is available.

For high-volume binary offline work, the recommended practical pattern is:

1. Run **Mistral Small 3.2 24B** on the full corpus.
2. Review the INCONCLUSIVE rate.
3. If the INCONCLUSIVE queue is too large or too important, run only those items through **Qwen 3.6 27B Dense** or **Qwen 3.6 35B MoE**.
4. Send remaining unresolved items to human review.

This avoids using a slow model on every email while still giving ambiguous items a stronger second pass.

Privacy and long-term control

Choose **offline** when privacy, air-gapped processing, model availability, or repeatability of the overall workflow is a top priority.

Mistral Small 3.2 24B is especially strong here because it combines:

- local processing;
- deterministic behavior;
- no per-token provider cost;
- no cloud rate limits;
- high Test 1 F1;
- practical throughput on consumer 24–32 GB GPUs;
- continued access after cloud providers retire or replace hosted models.

Reproducibility

For strict rerun consistency, prefer deterministic retained models.

Good starting points:

- **Offline:** Mistral Small 3.2 24B, Ministral 3 14B, Llama 3.3 70B.
- **Cloud:** Grok 4.2 Non-Reasoning.

Treat these retained models as non-deterministic or potentially variable between runs:

- Claude 4.7 Opus.
- Gemini 3 Flash.
- Gemini 3.1 Flash-Lite.
- OpenAI GPT-5.4.
- Qwen 3.6 27B Dense.
- Qwen 3.6 35B MoE.
- Gemma 4 26B Think.

Even with deterministic models, archive the exact output from the production run. Do not rely on rerunning any model later as the matter record.

Language support

Use benchmark language results as a guide, then validate on your own corpus.

Language / language mix	Practical starting point
English-dominant, high-volume binary	Mistral Small 3.2 24B offline; Grok 4.2 Non-Reasoning cloud if cloud is allowed and coverage matters.
French-heavy	Mistral-family models are especially attractive because French is a co-primary language.
Korean	Gemini 3.1 Flash-Lite or Gemini 3 Flash in cloud; Qwen 3.6 27B Dense or Gemma 4 26B offline; Mistral Small is a strong faster offline alternative but should be validated.
Japanese / Chinese	Claude 4.7 Opus, GPT-5.4, or Grok 4.2 Non-Reasoning in cloud; Qwen 3.6 or Gemma 4 offline.
Arabic / Hindi	Claude 4.7 Opus or GPT-5.4 in cloud; Gemma 4 26B offline.
Broad multilingual / multi-category triage	Gemini 3.1 Flash-Lite in cloud; Qwen 3.6 27B Dense offline when speed is acceptable.

Do not assume English results predict every language. Validate the actual language mix, including forwarded text and attachments if those will be included.

5. Practical Model Shortlist

Model	Deployment	Best fit	Avoid as first choice when...
Mistral Small 3.2 24B	Offline	Default practical offline model for high-volume binary classification; strong speed/accuracy balance; deterministic; consumer 24–32 GB GPU class.	Corpus uses unsupported or weakly supported languages; broad multi-category taxonomy is central; no review path exists for INCONCLUSIVE items.
Qwen 3.6 27B Dense	Offline	Accuracy-first multilingual and multi-category work; second-pass resolver for Mistral INCONCLUSIVE results.	Full-corpus production throughput is required on a single 24–32 GB workstation.
Qwen 3.6 35B MoE	Offline	Faster Qwen-family compromise for multilingual and multi-category work.	Dense-model consistency matters more than speed, or hardware cannot support the model comfortably.
Gemma 4 26B Think	Offline	Balanced offline multilingual option, especially when Korean support matters.	High-speed production classification is the main priority.

Ministral 3 14B	Offline	Fastest raw offline triage and lower-end 16 GB VRAM hardware.	Fine-grained multi-category discrimination is required.
Llama 3.3 70B	Offline	English-dominant, accuracy-focused offline work on high-end hardware.	Running on 32 GB VRAM and production throughput matters; Korean-heavy work is central.
Gemini 3.1 Flash-Lite	Cloud	Lowest-cost cloud classification; broad multilingual and multi-category cloud triage.	Strict reproducibility, UK/Swiss Gemini residency, or provider data transfer is unacceptable.
Grok 4.2 Non-Reasoning	Cloud	Highest cloud accuracy/coverage and deterministic cloud classification.	Budget, data-transfer restrictions, or required region make it unsuitable.
Claude 4.7 Opus	Cloud / Bedrock	Premium analysis, translation, reasoning-heavy review, and selected high-value matters.	Routine large-volume classification where speed and cost dominate.
OpenAI GPT-5.4	Cloud	Teams already standardized on OpenAI, or analysis workflows where GPT-specific capabilities matter.	Cost-sensitive routine classification.

6. Models and Choices to Treat Carefully

Avoid choosing a model solely because it is newer, larger, or more expensive.

Treat the following choices carefully:

- **OpenAI GPT-5.5**: excluded because it refused to classify the benchmark emails.
- **Grok 4.3**: excluded because it was slower and operationally dominated by Grok 4.2 Non-Reasoning.
- **Mistral Large 3, Magistral 24B, Nemotron 3 33B, Qwen 2.5 / 3.5 variants, GPT-OSS 20B variants, Gemma 4 E4B variants**: excluded from the retained benchmark set for underperformance, high abstention, or lack of advantage.
- **Very large offline models on insufficient hardware**: may run, but slow CPU offload can make production throughput impractical.
- **Any cloud model pinned to a long-term workflow**: provider retirement or behavior changes can break repeatability. Archive outputs and keep the model as a swappable component.

7. Validation Before Production

Benchmark rankings are the best current guideline, not an exact permanent ordering. Some retained models are non-deterministic, provider behavior can change, and model-task fit matters.

Before a full production run:

1. Confirm whether data may leave the organization.
2. Confirm required region, provider terms, logging, retention, and data-use policy.
3. Choose the model for the actual task shape: binary classification, multi-category triage, translation, summarization, or extraction.
4. Validate the language mix and any attachment-text setting.
5. Run a representative sample, including likely positives, likely negatives, borderline items, major custodians, key time periods, folders, and relevant languages.
6. Review all Responsive and INCONCLUSIVE results in the sample.
7. Spot-check Unresponsive results, especially where missed evidence would be costly.
8. Measure the INCONCLUSIVE rate.
9. If Mistral Small produces too many INCONCLUSIVE results, test a second-pass Qwen workflow on that subset.
10. Preserve the prompt, category list, model, provider, region, settings, date/time, attachment setting, validation notes, logs, and exported AI results.

---

8. Bottom Line

For most Aid4Mail practitioners with suitable hardware, **Mistral Small 3.2 24B is the practical offline anchor model** for high-volume binary classification. It is fast, accurate, deterministic, private, free of token charges, free of cloud rate limits, and not vulnerable to cloud-model retirement.

Use **Qwen 3.6 27B Dense** or **Qwen 3.6 35B MoE** as a targeted second pass when Mistral's INCONCLUSIVE queue needs further resolution, especially in multilingual or multi-category matters.

Use cloud models when you need low-friction setup, broad cloud language support, enterprise governance, very large context windows, or a specific cloud model's strengths. For cloud classification, the practical defaults are **Gemini 3.1 Flash-Lite** for lowest cost and multilingual breadth, and **Grok 4.2 Non-Reasoning** for highest cloud accuracy and decision coverage.

Date of publication: May 15, 2026.