Aid4Mail AI Provider and Model Selection Guide

User Guide

Fookes Software Ltd
Charmey, Switzerland
www.aid4mail.com

Table of Contents

Aid4Mail AI Provider and Model Selection Guide

Introduction

Aid4Mail supports a wide range of AI providers and models for email filtering, classification, and analysis. Choosing the right combination can significantly impact accuracy, speed, cost, and compliance. This guide helps you make that decision, whether you're a solo investigator on a budget or an enterprise team with strict data residency requirements.

This guide covers **what to choose and why**. For step-by-step API configuration, JSON editing, and advanced setup, refer to the Aid4Mail AI Provider and Model Configuration Guide technical guide. For full feature documentation, refer to the Aid4Mail AI Email Review Workflow Guide.

The model rankings in this guide are drawn from the Aid4Mail AI Classification Benchmark Report (May 2026). That program evaluated up to 40 AI model configurations across four accuracy tests (binary responsiveness in English, multi-category discrimination in English and two other Western languages, binary classification in Korean, multi-category discrimination in Korean), a large-scale Production Pilot on 34,097 Podesta emails, and a separate throughput and cost test (Test 5) on 1,083 Podesta emails averaging 81 KB each. Eleven model families—five cloud and six offline—were retained for the final benchmark set. Where deployment-specific behavior matters, Gemini 3.1 Flash-Lite is reported in two rows: Google AI Studio and Gemini Enterprise Agent Platform. The remainder were excluded because they were superseded by a newer or stronger sibling, underperformed, were unusable for forensic content (one cloud model refused to classify the corpus), or offered no advantage over a faster or cheaper alternative.

---

1. The Three Provider Categories

Aid4Mail supports three categories of AI provider, each suited to different needs.

1.1. Mainstream Cloud Providers

These are direct API services from AI companies. You create an account, get an API key, and start processing.

Provider	Models	Pricing Basis
Anthropic	Claude 4.7 Opus	Per token
Google	Gemini 3 Flash (preview), Gemini 3.1 Flash-Lite	Per token
OpenAI	GPT-5.4	Per token
xAI	Grok 4.2 Non-Reasoning	Per token

Best for: Getting started quickly, small projects with under 10,000 emails, teams comfortable with cloud processing.

Limitations: Shared rate limits can throttle sustained batch processing. Data is sent to the provider's servers—review their privacy policies before processing sensitive material.

1.2. Enterprise Cloud Platforms

These platforms host the same (or similar) models as mainstream providers, but add enterprise controls: predictable quotas, regional deployment, compliance certifications, and centralized billing. For marketing and regional-planning purposes, Google Vertex AI and Gemini Enterprise Agent Platform should be treated as one Google Cloud platform family rather than as separate alternatives.

Platform	Available Models	Key Advantage
Amazon Bedrock	Claude 4.7 Opus	Regional deployment across Americas, Europe, Asia-Pacific; reservable throughput
Google Vertex AI / Gemini Enterprise Agent Platform	Gemini, Claude	OAuth 2.0 authentication; Google Cloud governance; `us` and `eu` Gemini 3.1 Flash-Lite deployments
Microsoft Foundry	GPT-5.4, Grok 4.2	Multi-model access through a single account

Best for: Large-scale batch processing, organizations with data residency requirements (GDPR, PIPA, nFADP), teams needing predictable throughput.

How they differ from mainstream: Enterprise platforms don't eliminate rate limits—they replace shared, consumer-grade throttling with explicit quotas and deployment isolation better suited to sustained processing. They also provide enforceable data-residency guarantees, stronger contractual data-handling terms, and tenancy arrangements that are materially easier to defend to regulators and opposing counsel. The incremental pricing premium is small relative to the defensibility benefit.

1.3. Offline (Local) Providers

Run AI models on your own hardware. No data ever leaves your network.

Tool	Interface	How It Works
Ollama	Command line	Lightweight; automatic model management; preferred for Aid4Mail
LM Studio	Desktop GUI	Visual model browser and management

Best for: Air-gapped environments, classified investigations, organizations with strict data sovereignty requirements, eliminating recurring token costs.

Trade-offs: Requires local GPU hardware. The largest offline models (Llama 3.3 70B) require substantial VRAM (≥80 GB recommended) for full-speed operation. The mid-tier offline models retained in the benchmark (Mistral Small 3.2 24B, Ministral 3 14B, Gemma 4 26B Think, Qwen 3.6 27B Dense, Qwen 3.6 35B MoE) compete strongly with leading cloud models on accuracy while running comfortably on a single 16–32 GB GPU. Mistral Small 3.2 24B achieved the highest Test 1 F1 in the benchmark (99.6% on decided emails), and Llama 3.3 70B plus Qwen 3.6 35B MoE tied for second among offline models at 99.2% with near-complete decision coverage, so offline deployment is no longer a quality compromise for organizations with appropriate hardware.

---

2. Deciding Which Provider Category to Use

Start by asking these three questions:

Can your email data leave your network?

If **no**—due to regulation, policy, or the sensitivity of the investigation—your only option is **offline processing** with Ollama or LM Studio. Skip to Section 4.

If **yes, but with conditions** (such as data residency in a specific region), use an **enterprise platform** that deploys in your required region. See the regional availability table in Section 3.3.

If **yes, without restrictions**, all three categories are available. Continue below.

How many emails will you process?

Mainstream cloud providers are simple and effective, even for large batches. Rate limits are unlikely to be an issue, except for the following:

- **Claude models via the Anthropic API under Tier 1**, which imposes a limit of 30,000 input tokens per minute. In practice, this causes Aid4Mail to pause after every two or three emails, making processing unusably slow. Upgrade to Tier 2 or use Amazon Bedrock to avoid this limitation.
- **Gemini Flash models via Google AI Studio**, which are rate-limited on the free and low tiers. For higher volumes (10,000 or more emails), use Google Vertex AI / Gemini Enterprise Agent Platform instead.

For even more predictable throughput, enterprise platforms such as Amazon Bedrock, Google Vertex AI / Gemini Enterprise Agent Platform, and Microsoft Foundry provide explicit quotas and deployment isolation suited to sustained batch processing. Alternatively, offline processing eliminates rate limits entirely, at the cost of hardware investment.

What is your budget model?

Pay-per-use works well for occasional or moderate processing. Cloud providers charge per token, with costs ranging from $0.25 to $5.00 per million input tokens depending on the

model. At Test 5 payload sizes (~81 KB per email on average), this translated to roughly $42 per 100,000 emails for the cheapest viable cloud deployment (Gemini 3.1 Flash-Lite via Google AI Studio), roughly $43 per 100,000 emails for the faster Gemini Enterprise Agent Platform deployment, and around $1,275 per 100,000 emails for Claude 4.7 Opus (low effort). The cloud cost range is therefore about 30× from the cheapest retained cloud deployment to the most expensive.

Fixed investment suits teams with ongoing, high-volume workloads. Offline hardware (starting around $3,000) eliminates per-token costs after the initial purchase. Against Claude 4.7 Opus pricing, a single high-spec workstation reaches break-even after roughly a quarter-million emails processed; against the cheapest viable cloud model (Gemini 3.1 Flash-Lite at $0.25 per million input tokens), break-even shifts out to several million emails, and the case for offline rests primarily on data-residency and rate-limit grounds rather than cost.

3. Choosing an AI Model

Once you've selected a provider category, you need to pick a specific model. The four factors that matter most are accuracy, speed, cost, and language support.

3.1. Accuracy Rankings

The figures in this section are drawn from the May 2026 Aid4Mail AI Classification Benchmark. Four accuracy tests form the core of the benchmark:

- **Test 1**—2,000 emails (1,880 Podesta plus 120 synthetic) at 6% responsive prevalence; binary insider-threat / data-exfiltration classification.
- **Test 2**—200 synthetic emails across five misconduct categories plus clean/inconclusive, with a French and Spanish subset.
- **Test 3**—120 synthetic Korean emails; binary insider-threat classification.
- **Test 4**—150 synthetic Korean emails; multi-category misconduct classification.

Two complementary metrics are reported for Test 1. **F1** (the harmonic mean of precision and recall) is computed over the items the model classified as Responsive or Unresponsive; INCONCLUSIVE and no-result items are tracked separately via the **decision rate** rather than counted as errors. This aligns with the textbook definitions used in the TREC Legal Track and TAR literature. **Automation Yield (AY)** reports the share of the full 2,000-email corpus that the model resolved with a correct, committed binary decision—i.e., (TP + TN) ÷ Total. Where F1 measures the quality of the decisions the model committed to, AY measures how much of the corpus the automated stage fully handled without deferring to INCONCLUSIVE or Error. A model can lead on F1 yet trail on AY if it abstains often, and vice versa. AY is useful when the workflow cost of a manual review queue is non-trivial. AY is currently reported for Test 1 only.

Where deployment-specific behavior matters, Gemini 3.1 Flash-Lite is shown as two rows. The Google AI Studio and Gemini Enterprise Agent Platform deployments produced identical Test 1 F1 and identical Tests 2–4 accuracy, but differed in Automation Yield, speed, throughput, pricing, and regional availability.

Summary Results—All Retained Model Families and Deployment Rows

Model	Deployment	Test 1 F1	Test 1 AY	Test 2 Acc.	Test 3 Acc. (KR)	Test 4 Acc. (KR)	Input Cost/1M Tokens
Mistral Small 3.2 24B	Offline	99.6%	98.65%	97.5%	97.5%	97.3%	—
Llama 3.3 70B	Offline	99.2%	99.85%	98.5%	95.8%	93.3%	—
Grok 4.2 Non-Reasoning	Cloud	99.2%	99.85%	99.5%	99.2%	96.0%	$1.25
Qwen 3.6 35B MoE	Offline	99.2%	99.80%	99.5%	99.2%	98.0%	—
Qwen 3.6 27B Dense	Offline	98.8%	99.80%	100.0%	100.0%	100.0%	—
OpenAI GPT-5.4	Cloud	97.6%	98.70%	100.0%	98.3%	98.0%	$2.50
Claude 4.7 Opus (low)	Cloud	97.2%	99.65%	100.0%	99.2%	100.0%	$5.00
Gemini 3.1 Flash-Lite (Agent Platform)	Cloud	96.0%	99.50%	100.0%	100.0%	100.0%	$0.275
Gemini 3.1 Flash-Lite	Cloud	96.0%	99.45%	100.0%	100.0%	100.0%	$0.25
Gemma 4 26B Think	Offline	95.2%	99.30%	99.5%	100.0%	99.3%	—
Ministral 3 14B	Offline	94.9%	99.30%	93.0%	98.3%	91.3%	—
Gemini 3 Flash (preview)	Cloud	94.5%	99.25%	100.0%	100.0%	100.0%	$0.50

Important findings from Test 1. Every retained model achieved 99% or higher recall—meaning virtually no responsive emails are missed. Differences in F1 are driven almost entirely by precision (false positives), not recall. When the full corpus is taken into account, **Llama 3.3 70B and Grok 4.2 Non-Reasoning are tied for the top Automation Yield (99.85%)**: Llama's 99.2% F1 was computed over 1,999 decided emails (one routed to INCONCLUSIVE—a 99.95% decision rate), and Grok 4.2 Non-Reasoning matched it at the corpus level despite being a cloud model. The two Qwen 3.6 variants are close behind at 99.80%. Mistral Small 3.2 24B's benchmark-leading F1 of 99.6% was computed over 1,974 decided emails (26 routed to INCONCLUSIVE), which places it at the bottom of the retained

models on AY (98.65%). Neither result is wrong—they describe different workflows. Choose by F1 when you trust the model's abstentions and want only high-confidence decisions; choose by AY when every INCONCLUSIVE item costs reviewer time. **Grok 4.2 Non-Reasoning leads the cloud tier on both F1 (99.2%) and AY (99.85%)**, with Claude 4.7 Opus (low) close behind on cloud AY at 99.65% and Gemini 3.1 Flash-Lite at 99.45%–99.50%, depending on deployment.

Consistency across tests. The most consistent performer across English and Korean, binary and multi-category, is **Qwen 3.6 27B Dense**, with a worst-to-best spread of just 1.2 percentage points across all four tests. Qwen 3.6 35B MoE follows at 1.5 points, then Mistral Small 3.2 24B at 2.3 points, OpenAI GPT-5.4 at 2.4 points, Claude 4.7 Opus (low) at 2.8 points, and Grok 4.2 Non-Reasoning at 3.5 points. Gemini 3.1 Flash-Lite holds 100.0% on Tests 2, 3, and 4 and 96.0% on Test 1, giving it a 4.0-point spread in both deployments; it remains the strongest low-cost cloud pick for organizations whose workflows span multiple prompt shapes.

Models excluded from the final set. The benchmark evaluated 40 model configurations in total; 29 were excluded. Notable exclusions:

- **Superseded by newer or stronger siblings:** Claude Haiku 4.5, Claude Opus 4.5, Claude Opus 4.6, Claude Sonnet 4.6, Gemini 2.5 Flash, OpenAI GPT-5.2, Gemma 3 27B, Gemma 4 26B NoThink, Grok 4.1 Fast, Grok 4.1 Fast+Reasoning.
- **Excluded on low accuracy or high abstention:** Mistral Large 3, Magistral 24B, Nemotron 3 33B, Qwen 2.5 (14B and 32B), Qwen 3.5 9B, Qwen 3.5 27B (Think and NoThink variants), GPT-OSS 20B (Low and High), Gemma 4 E4B (Think and NoThink), and **OpenAI GPT-5.5** (refused to classify the benchmark emails, returning *"This content was flagged for possible cybersecurity risk"*—the model is unusable for forensic email work regardless of its underlying capability).
- **Excluded because a smaller offline model delivered equal or better accuracy at higher throughput and a far smaller VRAM footprint:** Gemma 4 31B (Think and NoThink), GPT-OSS 120B (Low and High).
- **Operational outliers worth flagging:** Grok 4.3 was by far the slowest cloud model tested—4 h 55 m on Test 1 versus 24 m for Grok 4.2 Non-Reasoning—and posted lower Test 1 F1 (96.3%) than its non-reasoning sibling. It is dominated on every operational dimension by Grok 4.2 Non-Reasoning.

Multi-category classification (Test 2 highlights). Frontier performance on the five-category misconduct task is essentially saturated at the top:

- **100% accuracy:** Gemini 3.1 Flash-Lite (Agent Platform), Gemini 3.1 Flash-Lite, Gemini 3 Flash (preview), OpenAI GPT-5.4, Claude 4.7 Opus, Qwen 3.6 27B Dense.
- **99.5% accuracy:** Grok 4.2 Non-Reasoning, Gemma 4 26B Think, Qwen 3.6 35B MoE.
- **98.5% accuracy:** Llama 3.3 70B.
- **97.5% accuracy:** Mistral Small 3.2 24B.
- **93.0% accuracy:** Ministral 3 14B—the seven-category prompt exceeds its discrimination capacity.

Four entries—three distinct model families—scored a perfect 100.0% on Tests 2, 3, and 4: Qwen 3.6 27B Dense (offline), Gemini 3 Flash (preview), and both deployments of Gemini 3.1 Flash-Lite. Qwen 3.6 27B Dense remains the only offline model in this perfect-Tests-2/3/4 group. Several models show divergent behavior between binary and multi-category work: Gemini 3 Flash (preview) scores 100% on Tests 2, 3, and 4 but sits at the bottom of the retained Test 1 F1 ranking (94.5%). Models should be evaluated on the prompt shape that matches the actual workflow, not on a single aggregate score.

3.2. Speed and Throughput

Processing speed matters for large jobs. The table below shows emails-per-second and estimated unattended 62-hour weekend throughput (Friday 18:00 to Monday 08:00) using **Test 5** as the reference where available. Test 5 used a 1,083-email Podesta corpus with no upper size filter and an average message size of ~81 KB—more representative of real-world payloads than Test 1, which was size-filtered to 1–68 KB so that the full model lineup could be evaluated within a reasonable total runtime. Test 1 speed figures are correspondingly optimistic for production use; where a model was not in Test 5, the table footnotes mark the figure accordingly.

Commercial Models

Model	Speed (emails/s)	Weekend Throughput (est.)	Cost / 100K emails	Input Cost/1M Tokens
Gemini 3.1 Flash-Lite (Agent Platform)	1.72	~384,000	~$43	$0.275
Gemini 3.1 Flash-Lite	1.30	~289,000	~$42	$0.25
Grok 4.2 Non-Reasoning[1]	1.35	~300,000	~$192	$1.25
Gemini 3 Flash (preview)	1.22	~272,000	~$119	$0.50
OpenAI GPT-5.4	0.98	~219,000	~$383	$2.50
Claude 4.7 Opus (low)[1]	0.47	~104,000	~$1,275	$5.00

[1] Test 1 figures (corpus filtered to 1–68 KB messages); Claude 4.7 Opus and Grok 4.2 Non-Reasoning were not part of the Test 5 throughput run. Production-scale throughput on heavier mailboxes (≥80 KB average) is expected to be similar to or slightly below the Test 1 numbers shown.

Note on Grok 4.2 Non-Reasoning. This is the cloud accuracy leader (Test 1 F1 = 99.2%, top-tier AY = 99.85%) and the natural successor to Grok 4.1 Fast in high-accuracy cloud workflows. On Test 1, it is the third-fastest retained cloud deployment at 1.35 emails/s, behind Gemini 3.1 Flash-Lite on the Agent Platform (1.72 emails/s) and Gemini 3.1 Flash-Lite via Google AI Studio (1.40 emails/s). Per-token pricing is higher than Grok 4.1 Fast was ($1.25 vs. $0.20 / 1M input tokens), but Grok 4.2's minimal output volume per email (~6 output tokens, comparable to Grok 4.1 Fast and roughly 50× lower than the Grok 4.1 Fast+Reasoning variant at ~296 output tokens per email) keeps total cost per 100K emails near $192—still substantially below GPT-5.4 ($383) and Claude 4.7 Opus ($1,275).

Note on Claude 4.7 Opus (low effort). Among retained cloud models, Claude 4.7 Opus is the slowest and by far the most expensive (1 h 11 m for 2,000 emails on Test 1; ~$1,275 per 100K emails on the same corpus). Its F1 (97.2%) is competitive with GPT-5.4 (97.6%) and below Grok 4.2 Non-Reasoning (99.2%), and its perfect score on the Korean multi-category test (Test 4 = 100%) is a clear strength for multilingual investigations. The cost–performance trade-off is unfavorable for routine classification; reserve Opus for workflows that benefit from its broader reasoning, summarization, and nuanced multilingual analysis.

Note on Gemini 3 Flash (preview) vs. Gemini 3.1 Flash-Lite. Gemini 3.1 Flash-Lite is now generally available and has two benchmarked deployments. Both deployments scored 96.0% F1 on Test 1 and 100.0% on Tests 2, 3, and 4. The Agent Platform deployment is the fastest cloud deployment in the benchmark at 1.72 emails/s on Test 5, finishing the Test 5 corpus about 25% sooner than Google AI Studio, or at about 33% higher emails/s. The

Google AI Studio deployment is the cheapest retained cloud option at roughly $42 per 100K emails. Gemini 3 Flash (preview) remains strong for multi-category and multilingual work, but it costs more than Flash-Lite and has lower Test 1 F1 (94.5%).

Offline Models

Model	Speed (emails/s)	Weekend Throughput (est.)	Min. VRAM
Ministral 3 14B	3.94	~879,000	16 GB
Mistral Small 3.2 24B	2.75	~614,000	24 GB
Gemma 4 26B Think	0.20	~45,500	24 GB
Qwen 3.6 35B MoE[1]	0.16	~34,600	32 GB
Llama 3.3 70B	0.14	~31,300	80 GB (or 32 GB with CPU offload, much slower)
Qwen 3.6 27B Dense[1]	0.08	~18,700	24 GB

[1] Test 1 figures; the two Qwen 3.6 variants were not part of the Test 5 throughput run. Both are reasoning models that produce substantial chain-of-thought output (≈670 tokens/email for the 27B Dense, ≈860 tokens/email for the 35B MoE), so the slow speeds reflect output volume rather than VRAM pressure. The MoE variant activates only ~3B parameters per token despite its 35B total, which is why it runs roughly twice as fast as the smaller 27B Dense at the same quantization.

Note on offline payload sensitivity. Offline throughput is sensitive to payload size in principle, but on this benchmark the prompt and context-window settings effectively bounded per-email input-token volumes—Test 5 inputs averaged only ~2–4% more tokens per email than Test 1 for the retained offline models present in both runs (Mistral Small 3.2 24B, Ministral 3 14B, Gemma 4 26B Think, Llama 3.3 70B), and offline emails/s on Test 5 was comparable to or slightly above Test 1 for those models. Real-world mailboxes with materially larger payloads, or configurations using longer context windows, will exhibit lower offline emails/s than the figures shown. Cloud throughput, by contrast, is largely bound by API round-trip latency and is approximately payload-insensitive at the speeds tested.

Note on Llama 3.3 70B. The 0.14 emails/s figure reflects partial CPU offload on the test system's 32 GB RTX 5090. A Q4_K_M Llama 3.3 70B requires approximately 80 GB to remain fully GPU-resident. With 96+ GB of VRAM (e.g., NVIDIA A100/H100), throughput is expected to increase substantially; with 32 GB, the slow speed reflects RAM offload, not the model's native capability.

Note on the Qwen 3.6 family. Both variants are reasoning-capable Ollama builds. The 27B Dense model posted perfect 100% scores on Tests 2, 3, and 4—the best multilingual offline performance in the benchmark—and is now the most consistent retained model across Tests 1–4, but at very low throughput (0.08 emails/s on Test 1). The 35B MoE variant is roughly twice as fast and gives up only 0.5–2 percentage points on the multilingual tests. Choose the Dense build for accuracy-critical multilingual passes on small corpora; choose the MoE for higher-volume offline runs where its 1.5-point worst-to-best spread remains valuable.

Offline speeds shown are from testing on an AMD Ryzen 9 9950X3D with an NVIDIA RTX 5090 (32 GB VRAM) and 192 GB DDR5 RAM. All offline models were served via Ollama at Q4_K_M quantization with 32K context length.

3.3. Regional Availability

For organizations with data residency requirements, the table below shows where models are hosted on enterprise platforms.

Platform	Model	Americas	Europe	Asia-Pacific
Amazon Bedrock	Claude 4.7 Opus	Brazil, Canada, USA	France, Germany, Ireland, Italy, Spain, Sweden, Switzerland, UK	Australia, Japan, Korea
Google Vertex AI / Gemini Enterprise Agent Platform	Gemini 3 Flash (preview)	Canada, USA	Belgium, Finland, France, Germany, Italy, Netherlands, Poland, Spain, UK	Australia, Japan, Korea
Google Vertex AI / Gemini Enterprise Agent Platform	Gemini 3.1 Flash-Lite	`us` multi-region	`eu` multi-region (excludes UK and Switzerland)	—
Google Vertex AI / Gemini Enterprise Agent Platform	Claude 4.7 Opus	USA	Belgium	—
Microsoft Foundry	GPT-5.4	USA	Sweden	—
Microsoft Foundry	Grok 4.2 Non-Reasoning	USA	Sweden	—

Regional availability changes frequently—always verify with the platform provider before deployment. For Gemini 3.1 Flash-Lite specifically, Aid4Mail surfaces the `us` and `eu` deployments; the `eu` deployment is a multi-region endpoint and does not include the UK or Switzerland. Customers requiring Swiss or UK residency for Gemini 3.1 Flash-Lite have no enterprise Gemini option at this time. For Swiss or UK residency on a premium cloud model, Claude 4.7 Opus via Amazon Bedrock is available; otherwise, use an offline model to keep processing on-premises. The integration layer should treat the model as a swappable component rather than locking documentation, prompts, or workflows to a specific cloud version.

3.4. Multilingual Support

Not all models handle all languages equally. The benchmark's Tests 3 and 4 probed Korean-language handling directly (Korean is typologically distant from English and uses a non-Latin script, making it a challenging test case). Test 2 included five emails per misconduct theme in French and five in Spanish (25 per language overall).

Key findings from the Korean tests:

- On the focused binary task (Test 3), five deployment rows across four model families achieved 100%: **Gemini 3.1 Flash-Lite (Agent Platform)**, **Gemini 3.1 Flash-Lite**, **Gemini 3 Flash (preview)**, **Gemma 4 26B Think**, and **Qwen 3.6 27B Dense**.

- On the multi-category task (Test 4), five deployment rows achieved 100%: **Gemini 3.1 Flash-Lite (Agent Platform)**, **Gemini 3.1 Flash-Lite**, **Gemini 3 Flash (preview)**, **Claude 4.7 Opus (low)**, and **Qwen 3.6 27B Dense**. **Gemma 4 26B Think** scored 99.3%, the best Korean multi-category result among models that did not score 100%.
- **Mistral Small 3.2 24B** performed strongly on both Korean tests (97.5% and 97.3%).
- **Qwen 3.6 35B MoE** held up well on Korean (99.2% and 98.0%) at substantially higher speed than the 27B Dense.
- **Llama 3.3 70B** showed the weakest Korean binary result of any retained model (95.8% on Test 3) and a weak Korean multi-category result (93.3% on Test 4) despite its top-tier English F1—a reminder that English-language benchmarks do not always predict multilingual behavior.
- **Ministral 3 14B** struggled on Korean multi-category work (91.3% on Test 4), driven mainly by compliance-violation misclassification, while still posting a respectable 98.3% on the focused binary task.
- **Grok 4.2 Non-Reasoning** matched its predecessor on Korean binary (99.2%, tied with Grok 4.1 Fast) and improved on it for Korean multi-category (96.0% versus 93.3% for Grok 4.1 Fast), but still trails the leaders on the multi-category task.

For **French and Spanish** content, all retained models maintained at least 93% accuracy on the Test 2 multilingual subset, though the small per-language sample size limits granularity.

Language support summary:

Model	English	French	German	Spanish	Italian	Korean	Japanese	Chinese	Arabic
Claude 4.7 Opus	★	●	●	●	●	●	●	●	●
GPT-5.4	★	●	●	●	●	●	●	●	●
Grok 4.2 Non-Reasoning	★	●	●	●	●	●	●	●	●
Gemini 3 Flash (preview) / 3.1 Flash-Lite	●	●	●	●	●	●	●	●	●
Llama 3.3 70B	★	●	●	●	●	○	●	●	○
Gemma 4 26B Think	★	●	●	●	●	●	●	●	●
Mistral Small 3.2 24B	★	★	●	●	●	●	○	○	○
Ministral 3 14B	★	★	●	●	●	●	○	○	○
Qwen 3.6 27B Dense / 35B MoE	●	●	●	●	●	●	●	●	○

★ = Primary language; ● = Strong support; ○ = Basic support

Key takeaways:

- For **European languages** (English, French, German, Spanish, Italian), all retained cloud models and most offline models perform well. Mistral-family models (Mistral Small, Ministral) have French as a co-primary language, making them particularly suited to French-language datasets.
- For **Korean**, **Gemini 3.1 Flash-Lite** (either deployment) and **Gemini 3 Flash (preview)** all achieve perfect scores on both Tests 3 and 4; **Claude 4.7 Opus (low)** achieves perfect Test 4 but one Test 3 error. Offline, **Qwen 3.6 27B Dense** is perfect on both Korean tests, while **Gemma 4 26B Think** is perfect on Test 3 and 99.3% on Test 4. Mistral Small 3.2 24B is the best alternative for offline Korean work on modest hardware where speed matters.
- For **Japanese or Chinese**, Claude 4.7 Opus and GPT-5.4 provide strong commercial support. Gemma 4 26B Think and the Qwen 3.6 family are the recommended offline options.
- For **Arabic or Hindi**, Claude 4.7 Opus and GPT-5.4 provide strong commercial support. Gemma 4 26B Think is the recommended offline option.
- For **datasets spanning multiple language families** (e.g., English and Korean together), choose a commercial model with broad strong support, or consider running separate passes with models optimized for each language group.

3.5. Reproducibility

For most classification workflows, running the same prompt against the same email twice produces the same answer. Some models do not behave this way, which matters in forensic and eDiscovery contexts where the ability to reproduce a classification supports its defensibility.

The cause varies by model—internal sampling during reasoning, an API-level restriction on the temperature parameter, or both. The Aid4Mail AI Provider and Model Configuration Guide document explains the mechanisms in detail and marks all affected models with † in its directory tree.

Affected models in the retained benchmark set:

- **Cloud:** Claude 4.7 Opus, Gemini 3 Flash (preview), Gemini 3.1 Flash-Lite, OpenAI GPT-5.4
- **Offline:** Qwen 3.6 27B Dense, Qwen 3.6 35B MoE, Gemma 4 26B Think

Implications for model selection:

- **Variability is typically small.** Run-to-run differences appear in genuinely ambiguous emails near the responsive/unresponsive boundary, not in clear-cut cases.
- **If strict reproducibility is required**, the strongest deterministic alternatives in the retained set are **Grok 4.2 Non-Reasoning** (cloud), and **Mistral Small 3.2 24B**, **Llama 3.3 70B**, and **Ministral 3 14B** (offline). All accept temperature 0.0 and run without internal reasoning.
- **Archive the classification output.** For challenge defense, the output saved at classification time is the record of the decision. Re-runs of a non-deterministic model are not guaranteed to reproduce a contested classification.

3.6. Model Selection Decision Tree

Use this decision tree to narrow your choice:

```
START
  │
  ├─ Must data stay on-premises? ──── YES ──→ See Section 4 (Offline Models)
  │
  NO
  │
  ├─ Is legal defensibility the top priority?
  │   │
  │   YES ──→ Grok 4.2 Non-Reasoning (top cloud F1 and top cloud AY)
  │           or Gemini 3.1 Flash-Lite (perfect recall, strongest
  │           low-cost multilingual cloud profile)
  │           or Claude 4.7 Opus via Bedrock for enterprise governance
  │           and premium analysis capability
  │
  │   NO
  │   │
  │   ├─ Processing 100,000+ emails?
  │   │   │
  │   │   YES
  │   │   │
  │   │   ├─ Multi-category triage across themes?
  │   │   │   │
  │   │   │   YES ──→ Gemini 3.1 Flash-Lite via Google Vertex AI /
  │   │   │           Gemini Enterprise Agent Platform where eligible
  │   │   │           (100% on Tests 2, 3, and 4; lowest cloud cost;
  │   │   │            Agent Platform is fastest for US/EU deployments)
  │   │   │
  │   │   └─ Binary responsiveness?
  │   │       └──→ Grok 4.2 Non-Reasoning (F1 = 99.2%, AY = 99.85%,
  │   │            ~300,000 emails per weekend at ~$192 / 100K)
  │   │            or Gemini 3.1 Flash-Lite if budget dominates
  │   │            (~$42 / 100K, F1 = 96%)
  │   │
  │   │   NO
  │   │   │
  │   │   ├─ Budget is the primary constraint?
  │   │   │   │
  │   │   │   YES ──→ Gemini 3.1 Flash-Lite ($0.25/1M tokens, F1 = 96%)
  │   │   │
  │   │   │   NO ──→ Claude 4.7 Opus when premium analysis/reasoning
  │   │   │          capability or top-tier Korean multi-category
  │   │   │          accuracy justifies the price
```

3.7. Large-Scale Operational Validation

The headline performance figures in Sections 3.1 and 3.2 are derived from the 2,000-email Test 1 corpus and the 1,083-email Test 5 corpus. To verify that those figures hold at realistic production scale—and to surface any behaviors that only become apparent at volume—the benchmark program included a **Production Pilot** on the publicly available John Podesta

Emails corpus, a dataset widely used as a reference benchmark by forensics and eDiscovery practitioners.

Test design. A pre-filtered subset of 34,097 emails (3.62 GB), reduced from 50,887 originals using the Aid4Mail filter query `Type:Personal AND NOT Type:Duplicate`, was classified using a detailed FOIA-style political campaign prompt into three categories: Responsive, Unresponsive, and INCONCLUSIVE. The Production Pilot was conducted on March 2, 2026, before Gemini 3.1 Flash-Lite was released; it used **Gemini 2.5 Flash** (the then-current Google Flash model, since superseded) via Google Vertex AI (europe-west1) and **Mistral Small 3.2 24B** via Ollama. Each model processed the full corpus twice—once excluding attachment data and once including extracted document text. No pre-verified ground truth was established; results are reported in terms of classification distribution and partial manual review. A future run with the newer Gemini 3.1 Flash-Lite is possible, but the operational insights below remain useful.

Throughput estimates confirmed. Gemini 2.5 Flash processed 68,194 email classifications (both runs combined) in 10 hours 32 minutes at a total cost of $36.43, directly confirming a weekend throughput estimate of approximately 400,000 emails for under $220. Mistral Small 3.2 24B achieved a comparable ~410,000 emails per weekend at zero marginal cost once hardware is amortized.

Mistral Small 3.2 24B throughput is payload-dependent. On this corpus—where Mistral processed an average of approximately 1,677 input tokens per email across both runs—Mistral Small 3.2 24B achieved 2.07 emails/s without attachments and 1.66 emails/s with document attachments. Both context lengths tested (32K without attachments and 64K with) kept the model fully GPU-resident on the 32 GB VRAM test system; the speed difference between the two runs is attributable to the larger payload from included attachments, not KV cache overhead. Heavier business correspondence (larger average message size, attachment-rich mailboxes) will reduce throughput further.

Gemini 2.5 Flash throughput is insensitive to payload size. Gemini processed both runs at a stable 1.78–1.82 emails/s regardless of whether attachments were included, confirming that cloud-model throughput is constrained by API overhead per call rather than raw compute.

Both models agreed on approximately 99% of the corpus. Unresponsive classifications were nearly identical—33,742–33,754 for Gemini and 33,808–33,819 for Mistral. The divergence was confined entirely to the gray zone: Gemini classified 136–147 emails as Responsive and 207–208 as INCONCLUSIVE, while Mistral classified only 11 emails as Responsive and 267–278 as INCONCLUSIVE. The models disagreed not on whether to flag content, but on how to resolve uncertainty—Gemini tended toward Responsive, Mistral toward INCONCLUSIVE.

Partial manual review suggests Mistral was more precise, Gemini more thorough. The 11 Responsive emails identified by Mistral were confirmed as correctly classified in the reviewed sample, though Mistral also missed some genuinely responsive emails. A significant proportion of Gemini's Responsive classifications were judged borderline or false positives in the reviewed sample. For investigations where false positives are costly—for example, where Responsive emails trigger intensive manual review—Mistral's conservative behavior may be preferable. Where under-inclusion carries greater risk, Gemini's broader recall may be more appropriate.

Prompt complexity amplifies model differences. The prompt used for this scale test was substantially more nuanced than typical binary prompts, specifying multiple conditional criteria, exclusions for forwarded content, and an explicit default-to-Unresponsive instruction. The significant divergence in Responsive counts suggests that prompt complexity is itself a variable in model behavior—not just a parameter to be optimized. When using highly specific classification criteria for sensitive review tasks, test across at least two models on a representative sample before large-scale processing, and consider whether your priority is recall (flagging more for manual review) or precision (minimizing false positives).

Attachment inclusion produced minimal benefit for this corpus. Enabling document attachment extraction increased the Responsive count by 11 emails (8%) for Gemini and by zero for Mistral, while increasing input token volume by approximately 32% for both models and adding 7 minutes (Gemini) to 67 minutes (Mistral) of processing time. The incremental benefit did not justify the additional cost and processing time for this corpus. This finding is specific to this dataset and prompt, where key content was concentrated in email bodies; for investigations where critical evidence is more likely to be embedded in attached documents, the calculus may differ. Note that Aid4Mail always includes each attachment's filename and extension in the payload regardless of whether full text extraction is enabled—an attachment named `client_list_2026.xlsx` or `salesforce_export.csv` is already classifiable on name alone.

---

4. Offline Models: A Complete Guide

Running models locally gives you complete data privacy but requires careful planning around hardware, model selection, and configuration. This section covers everything you need to know.

4.1. How Offline AI Works in Aid4Mail

Aid4Mail connects to a local inference server—either **Ollama** (command-line) or **LM Studio** (desktop GUI)—running on your machine or network. The server loads an AI model into your GPU's memory and processes requests locally. No data is sent to external servers.

Aid4Mail ships with pre-built configurations for tested offline models. You don't need to edit any JSON files—just install the inference tool, download a model, and select it in Aid4Mail.

4.2. Understanding Parameter Size, VRAM, and Performance

Every AI model has a **parameter count** (measured in billions—"B"). Larger models are generally more accurate but require more powerful hardware and run slower. Mixture-of-Experts (MoE) models are an important exception: they hold a large total parameter count but activate only a small subset per token, so they can run substantially faster than a dense model of the same total size.

The critical resource is VRAM—the memory on your GPU. When a model and its working memory (called the "KV cache") fit entirely within VRAM, inference runs at full GPU speed. When they don't fit, the overflow spills to system RAM, and processing can slow down dramatically—often by 2× or more.

Here's how the retained offline models compare, based on the May 2026 benchmark:

Model	Parameters	Min. VRAM (Q4_K_M)	Test 1 F1	Test 5 Speed (emails/s)
Ministral 3 14B	14 B (dense)	16 GB	94.9%	3.94
Mistral Small 3.2 24B	24 B (dense)	24 GB	99.6% (on 1,974 decided)	2.75
Qwen 3.6 35B MoE	35 B total / ~3 B active	32 GB	99.2%	0.16[1]
Llama 3.3 70B	70 B (dense)	80 GB	99.2%	0.14 (32 GB w/ offload)
Qwen 3.6 27B Dense	27 B (dense)	24 GB	98.8%	0.08[1]
Gemma 4 26B Think	26 B (dense)	24 GB	95.2%	0.20

[1] Test 1 figures (size-filtered corpus); the two Qwen 3.6 variants were not part of the Test 5 throughput run.

Key observations:

- **Mistral Small 3.2 24B** achieves the highest Test 1 F1 of any model in the benchmark on decided emails (99.6%), combined with strong offline throughput (2.75 emails/s on Test 5). It does abstain more than the other offline models (26 INCONCLUSIVE responses on Test 1 versus at most a handful for the others), so workflows that cannot accommodate a small human-review queue should prefer a model with a higher decision rate.
- **Llama 3.3 70B** ties for second overall on F1 (99.2% across 1,999 decided emails, with only one INCONCLUSIVE routing) and ties with cloud-side Grok 4.2 Non-Reasoning for the top Automation Yield in the benchmark (99.85%). It is the strongest English-language choice when your hardware budget can accommodate ≥80 GB VRAM. On a 32 GB card with system-RAM offload, throughput drops to ~0.14 emails/s.
- **Qwen 3.6 27B Dense** is the most accurate offline model across the multilingual tests, posting perfect 100% scores on Tests 2, 3, and 4 and a Test 1 F1 of 98.8%. It is also the most consistent retained model across all four tests, with a 1.2-point worst-to-best spread. The trade-off is throughput: at 0.08 emails/s on Test 1, it is the slowest retained model in the benchmark, driven mostly by reasoning-output volume rather than raw VRAM pressure.
- **Qwen 3.6 35B MoE** (35B total, ~3B active) runs roughly twice as fast as the 27B Dense (0.16 emails/s on Test 1) at the same quantization, gives up only 0.5–2 percentage points across the multilingual tests, and ties for second on Test 1 F1 at 99.2%. It is the best general-purpose offline pick when consistency and balanced throughput matter more than peak multilingual accuracy.
- **Gemma 4 26B Think** is the best-balanced *fast-loading* offline model across all four accuracy tests—near-perfect on Tests 2, 3, and 4, including 100% on Korean binary and 99.3% on Korean multi-category, the best Test 4 result among models that did not score 100%. Slower than Mistral Small on the reference hardware (0.20 emails/s).
- **Ministral 3 14B** is the fastest model in the benchmark at any price (3.94 emails/s on Test 5), runs on 16 GB VRAM (RTX 5080 GPU), and is one of four retained offline models with native 256K context support, although 32K remains the practical setting

on 16 GB hardware. Its binary F1 is competitive (94.9%), but multi-category accuracy drops noticeably (93.0% on Test 2, 91.3% on Test 4), so it is best used for focused binary passes rather than wide multi-category triage.

- **Accuracy does not scale linearly with parameter count.** The 14-billion-parameter Ministral runs roughly even with the 26-billion-parameter Gemma 4 on Test 1 F1 (94.9% vs. 95.2%), the 24-billion-parameter Mistral Small leads the benchmark overall, and the 35-billion-MoE Qwen 3.6 outpaces the dense 27-billion Qwen 3.6 sibling without losing meaningful accuracy.

4.3. Quantization: Why Q4_K_M Is Recommended

Downloaded models come in different precision levels called **quantization**. Lower precision reduces the model size and speeds up processing, with minimal impact on classification accuracy.

Aid4Mail recommends **Q4_K_M** quantization for all offline models. Testing showed that higher-precision variants (Q8_0, FP16) significantly reduced processing speed without any meaningful improvement in classification accuracy for email tasks.

When downloading models with Ollama, the default download is typically Q4_K_M—you don’t need to do anything special.

4.4. Context Length: A Critical Performance Setting

Context length determines how much text the model can process in a single request. It directly affects both what fits (emails plus attachments) and how fast the model runs.

Why Context Length Matters

Every request allocates a block of GPU memory for the KV cache, proportional to the context length. A larger context length means:

- **Benefit:** Larger emails (especially those with attachment text) can be analyzed without truncation.
- **Cost:** More VRAM is consumed, potentially pushing the model out of GPU memory and causing severe slowdowns.

In testing, reducing Mistral Small 3.2 24B from 128K to 64K context produced a **2.4× speed improvement** with zero email truncation and no change in accuracy. The May 2026 benchmark used 32K context across all offline models, and that setting handled the Test 1 corpus without truncation issues.

Recommended Context Lengths by VRAM

The table below shows the largest safe context length for each model class at Q4_K_M quantization, given your GPU’s VRAM.

VRAM	14B	20–27B	35B MoE	70B
16 GB	32K	—	—	—
24 GB	64K	32K	—	—
32 GB	128K	64K	32K	—

40 GB	256K*	64K	64K	—
48 GB	256K*	128K	64K	—
80 GB	256K*	256K*	128K	64K
96 GB	256K*	256K*	256K*	128K

*256K applies only to models with native 256K architecture, including Gemma 4 26B, Ministral 3 14B, Qwen 3.6 27B Dense, and Qwen 3.6 35B MoE. Use 256K only when large attachment text requires it; 32K–64K remains the practical default for most email-classification workloads.

Notes:

- Mistral Small 3.2 24B and Llama 3.3 70B remain 128K-native offline models.
- The 35B MoE column reflects Qwen 3.6 35B-A3B, where the full 35B parameters must reside in VRAM but only ~3B parameters are active per token. Memory footprint is set by the full model size, not by the active parameter count.
- Setting a high context length reserves KV cache memory upfront, even for short prompts. This means oversized context lengths waste VRAM without benefit on shorter emails.
- Performance drops sharply at very high context sizes due to KV cache growth. For batch-style email classification workloads in Aid4Mail, **32K–64K is typically the best balance of speed and VRAM efficiency**, unless your emails regularly include large attachments that require a larger context window.
- If you experience unexpectedly slow processing, try reducing the context length—this is often the single most effective tuning adjustment.
- "—" indicates that the model cannot be kept fully GPU-resident at any viable context length on that VRAM tier, meaning inference will partially offload to CPU and throughput will be severely degraded. The model may still run, but not at production-viable speeds.

Practical Guidance

- **Start with 32K** for forensic and personal email corpora with average payloads under ~2,000 tokens. The May 2026 benchmark used this setting across all retained offline models without truncation issues. Production-Pilot testing on 34,097 emails of real-world correspondence at 32K produced only one truncated email—and that email was independently corrupted—confirming this is a practical production setting for short-to-medium email corpora without large documents.
- **Use 64K** if your emails regularly include attachment text and your VRAM can accommodate it.
- **Use 128K or 256K only** if your emails regularly include very large attachments, your model natively supports the target context length, and your VRAM can accommodate it without pushing the model to system RAM. In practice, 256K requires at least 40 GB VRAM for the 14B row, 80 GB for the 20–27B row, and 96 GB for the 35B MoE row.
- **If processing feels slow**, reducing the context length is often the single most effective fix.
- Aid4Mail reports when email messages are truncated to fit the context window. If truncation rates exceed a few percent, increase the context length—but test the speed impact before committing.

4.5. Choosing an Offline Model

Use this decision tree based on your hardware:

```
What GPU do you have?
  │
  ├─ No dedicated GPU or < 16 GB VRAM
  │   └─ Offline AI is not recommended.
  │      Consider a cloud provider, or upgrade your hardware.
  │
  ├─ 16 GB VRAM (e.g., RTX 5080, RTX 4080)
  │   └─ Ministral 3 14B (primary recommendation)
  │      Test 1 F1 = 94.9%, fastest model in the benchmark
  │      (3.94 emails/s on Test 5), native 256K context support
  │      with sufficient VRAM. Use 32K on 16 GB hardware.
  │      Best binary classifier in this hardware tier.
  │      Note: weaker on multi-category prompts—use a different
  │      model for wide thematic taxonomies.
  │
  ├─ 24 GB VRAM (e.g., RTX 4090, RTX 5090 LE)
  │   ├─ Need top multilingual / multi-category accuracy:
  │   │   └─ Qwen 3.6 27B Dense (100% on Tests 2, 3, and 4;
  │   │      most consistent retained model at 1.2 pts spread)
  │   │      for small, accuracy-critical batches, OR Qwen 3.6
  │   │      35B MoE for higher throughput at near-equal accuracy
  │   │      (1.5 pts spread).
  │   │
  │   ├─ Need balanced multilingual performance with faster
  │   │  loading and faster cold-start than Qwen 3.6:
  │   │   └─ Gemma 4 26B Think
  │   │      Test 1 F1 = 95.2%, 99.5% on Test 2, 100% Korean binary,
  │   │      99.3% Korean multi-category. Native 256K context support
  │   │      requires far more VRAM; use 32K on 24 GB hardware.
  │   │
  │   └─ Need top binary F1 and speed:
  │       └─ Mistral Small 3.2 24B
  │          Benchmark-leading F1 = 99.6% (on 1,974 decided emails),
  │          2.75 emails/s on Test 5. Abstains slightly more than
  │          alternatives (26 INCONCLUSIVE on Test 1)—well suited
  │          to workflows that treat the three-label design as a
  │          feature rather than a defect.
  │
  ├─ 32 GB VRAM (e.g., RTX 5090)
  │   └─ Same recommendations as above, with headroom for
  │      larger context windows (64K). Llama 3.3 70B will run
  │      with CPU offload but at degraded speed (~0.14 emails/s).
  │
  └─ 80+ GB VRAM (e.g., dual A6000, A100, H100)
      └─ Llama 3.3 70B (primary recommendation for English work)
         Tied for second-highest F1 in the benchmark (99.2% at a
         99.95% decision rate, with near-perfect English
         precision) and tied for the top Automation Yield
         of any retained model (99.85%)—effectively no
         manual review queue. Throughput is materially
         better at this VRAM tier than the 32 GB offload
         scenario. For multilingual / Korean-heavy workloads,
         pair it with Qwen 3.6 27B Dense or Gemma 4 26B Think
         since Llama's Korean accuracy trails the leaders.
```

4.6. Reasoning Models in Offline Deployment

Reasoning models—those that produce chain-of-thought tokens before their final answer—have specific deployment constraints offline.

Ollama and LM Studio enforce structured output through GBNF (GGML Backus-Naur Form) grammar constraints at the token level, masking any token that does not conform to the declared JSON schema. For models that depend on a dedicated reasoning token to initiate chain-of-thought, this constraint sets the probability of the reasoning token to zero, effectively disabling the feature that makes these models competitive on harder prompts.

The practical consequence: every offline reasoning-capable model in Aid4Mail's configuration is run with **unstructured output** to remain viable. This includes Gemma 4 26B Think and the two Qwen 3.6 reasoning variants. Several non-reasoning models—including Mistral Small 3.2 24B and Ministral 3 14B—are also run unstructured because prior Aid4Mail testing found that the Mistral family in particular produces comparable or better classification under unstructured output and runs faster.

Output-format choice is a **per-model empirical question**, not a deployment default. Aid4Mail's shipped configurations encode the empirically best setting for each supported model.

4.7. Setting Up Ollama (Step by Step)

Ollama is the recommended local inference tool for Aid4Mail.

Step 1: Install Ollama

Download and install from ollama.com.

Step 2: Download a Model

Open a terminal or command prompt and run one of these commands:

Model	Command
Ministral 3 14B	`ollama pull ministral-3:14b`
Mistral Small 3.2 24B	`ollama pull mistral-small3.2:24b`
Gemma 4 26B Think	`ollama pull gemma4:26b`
Qwen 3.6 27B Dense	`ollama pull qwen3.6:27b`
Qwen 3.6 35B MoE	`ollama pull qwen3.6:35b-a3b`
Llama 3.3 70B	`ollama pull llama3.3:70b`

The download happens automatically on first pull. Large models may take some time.

Step 3: Start the Server

Ollama's server usually starts automatically after installation. If not, run:

```
ollama serve
```

The server runs at `http://127.0.0.1:11434` by default. You can verify it's running by visiting that URL in a browser or running `ollama list`.

Step 4: Configure Aid4Mail

1. Open Aid4Mail and go to **App Settings > AI**.
2. Under **Provider configurations**, locate **Ollama** and then click **Configure…**.
3. Tick the **Available** checkbox.
4. Enter the **Context Length** as a raw token count. Use the conversion table below for the shorthand 'k' values used by Ollama:

Context Setting	Aid4Mail Value
32k	32768
64k	65536
128k	131072
256k	262144

1. Verify the endpoint URL is `http://127.0.0.1:11434` (the default).
2. In **Project Settings > AI**, select your model (e.g., **Gemma 4 26B Think (Ollama)**) for the task you want to run (Filter, Classify, or Analyze).

That's it. Ollama and the model are ready to use.

4.8. Setting Up LM Studio (Step by Step)

LM Studio is an alternative for users who prefer a graphical interface.

Step 1: Install LM Studio

Download and install from lmstudio.ai.

Step 2: Download and Load a Model

1. Open LM Studio and go to the **Discover** tab.
2. Search for a recommended model (see the table in Section 4.5).
3. Click **Download**.
4. Once downloaded, go to the **Chat** tab and select the model to load it.

Step 3: Start the Server

1. Navigate to the **Developer** tab.
2. Click **Start Server**. The server runs at `http://127.0.0.1:1234` by default.

Step 4: Configure Aid4Mail

1. Open Aid4Mail and go to **App Settings > AI**.
2. Under **Provider configurations**, locate **LM Studio** and then click **Configure…**.
3. Tick the **Available** checkbox.
4. Enter the **Context Length** as a raw token count.
5. Verify the endpoint URL is `http://127.0.0.1:1234` (the default).
6. In **Project Settings > AI**, select **Local Model (LM Studio)** for your task.

Note: LM Studio serves whichever model is currently loaded. If you change the model in LM Studio, Aid4Mail automatically uses the new model on subsequent requests—no Aid4Mail configuration change is needed.

5. Example Scenarios

5.1. Small Forensic Firm, No Dedicated GPU

Situation: A two-person forensic firm processes 5,000–20,000 emails per case, working primarily with English-language email. Budget is limited, and the workstation has no dedicated GPU suitable for AI inference.

Recommendation:

- **Provider:** Mainstream cloud (direct API)
- **Model:** Gemini 3.1 Flash-Lite
- **Why:** The cheapest viable cloud model at Test 5 payload sizes (~$42 per 100,000 emails via Google AI Studio), with strong cloud F1 (96.0% on Test 1, perfect recall, 10 false positives across 2,000 emails) and the strongest low-cost cloud multilingual profile (100% on Tests 2, 3, and 4). At 1.30 emails/s on Test 5, a 20,000-email case finishes in about 4¼ hours of wall-clock time and costs roughly $8–$9 in API fees. Use the Agent Platform deployment when the project is eligible for the `us` or `eu` endpoint and speed matters.
- **Alternative for higher accuracy:** Grok 4.2 Non-Reasoning (Test 1 F1 = 99.2%, AY = 99.85%, ~$192 / 100K) when reducing false positives is worth the higher per-email price.

5.2. Enterprise with Strict Data Residency (EU)

Situation: A European law firm handles cross-border litigation involving multilingual correspondence. GDPR requires that email data remain within the EU during processing. Volume is high (100,000+ emails per matter).

Recommendation:

- **Provider:** Enterprise cloud platform
- **Primary model:** Claude 4.7 Opus via Amazon Bedrock (deployed in Germany, France, or Ireland) for accuracy-critical work with enterprise-grade governance—particularly when Korean or Japanese content is in scope (Test 4 = 100%).
- **High-volume alternative:** Gemini 3.1 Flash-Lite via Google Vertex AI / Gemini Enterprise Agent Platform in the `eu` multi-region deployment—100% on Tests 2, 3, and 4, the fastest retained cloud throughput when using the Agent Platform deployment (1.72 emails/s on Test 5), and roughly $43 per 100,000 emails. The `eu` deployment does not include the UK or Switzerland; use Claude 4.7 Opus via Amazon Bedrock or an offline model when those specific residency requirements apply.
- **Why enterprise:** Predictable quotas, EU-hosted infrastructure, compliance documentation, and data processing agreements that are materially easier to defend to regulators and opposing counsel.

5.3. Government Agency, Air-Gapped Environment

Situation: A government forensic unit analyzes English-language email in a classified environment with no internet connectivity. The workstation has an NVIDIA RTX 5090 (32 GB VRAM).

Recommendation:

- **Provider:** Ollama (offline)
- **Primary model:** Gemma 4 26B Think at 32K context length for multilingual / multi-category work, or Mistral Small 3.2 24B for highest-F1 binary classification.
- **Why:** Gemma 4 26B Think delivers near-perfect accuracy across Tests 2, 3, and 4 (including 100% and 99.3% on Korean binary and multi-category), making it the strongest fast-loading offline choice when thematic triage or multilingual content is in scope. Gemma 4 26B also supports a native 256K context window, although reaching that context length in practice requires substantially more VRAM than the 32 GB reference workstation. Mistral Small 3.2 24B tops the entire benchmark on Test 1 F1 (99.6% on decided emails) at much higher throughput (2.75 emails/s on Test 5), so it is better suited to high-volume binary work that can absorb a modest INCONCLUSIVE review queue.
- **Multilingual maximum-accuracy alternative:** Qwen 3.6 27B Dense for cases where the highest possible multilingual accuracy is required—it scored 100% on Tests 2, 3, and 4—accepting much lower throughput (0.08 emails/s on Test 1).
- **Highest-accuracy alternative for English:** If the agency can deploy 80+ GB VRAM (dual high-end GPUs or A100/H100), Llama 3.3 70B is tied for the second-highest F1 in the benchmark (99.2% at a 99.95% decision rate) and runs at materially better speeds than the 32 GB offload scenario.

5.4. International Investigation with Multilingual Email

Situation: A corporate investigation spans offices in the US, Germany, Korea, and Japan. The email collection includes all four languages in substantial volume. Accuracy is critical—false negatives could miss key evidence.

Recommendation:

- **Provider:** Enterprise cloud (Google Vertex AI / Gemini Enterprise Agent Platform or Amazon Bedrock)
- **Primary model:** Gemini 3.1 Flash-Lite—the strongest low-cost cloud multilingual performer in the benchmark, with perfect 100% results on Tests 2, 3, and 4 in both deployments. Use the Agent Platform deployment when the project is eligible for the `us` or `eu` endpoint and speed matters.
- **Secondary option:** Claude 4.7 Opus for analysis-heavy passes where reasoning, summarization, and a perfect Korean multi-category result (Test 4 = 100%) justify the premium pricing. It achieved one Test 3 error but remains strong across all tested languages.
- **Alternative approach:** For a two-pass strategy, use Gemini 3.1 Flash-Lite or Grok 4.2 Non-Reasoning for initial high-speed triage of the full collection, then run Claude 4.7 Opus on the flagged subset (Responsive plus Inconclusive) for final classification. This concentrates the slower model's effort on the items that matter and typically cuts slow-model token volume by 90%+ on corpora with prevalence in the 5%–15% range.

5.5. Cost-Sensitive eDiscovery, Very Large Volume

Situation: A litigation support company needs to classify 500,000 emails from a corporate merger review. Budget is the primary constraint, but accuracy must remain defensible. Emails are primarily in English.

Recommendation:

- **Provider:** Google AI Studio (direct API) or Google Vertex AI / Gemini Enterprise Agent Platform when enterprise governance or `us`/`eu` deployment is required
- **Model:** Gemini 3.1 Flash-Lite
- **Why:** At 1.30 emails/s on Test 5 payloads and roughly $42 per 100,000 emails through Google AI Studio, this model classifies the full 500,000-email corpus in under two unattended weekends for approximately $210 in API fees. The Agent Platform deployment is faster at 1.72 emails/s and costs roughly $43 per 100,000 emails. Test 1 F1 of 96.0% with perfect recall is at the upper end of the TAR 2.0 (CAL) envelope, making this defensible for first-pass review with manual quality checks on a sample. Automation Yield of 99.45%–99.50% means the unresolved-or-incorrect tail is about 2,500–2,750 items per 500,000 in the benchmark pattern, mostly concentrated in false positives and any INCONCLUSIVE results.
- **Accuracy alternative:** Grok 4.2 Non-Reasoning at 1.35 emails/s and ~$192 / 100K—higher F1 (99.2%) and top cloud AY (99.85%), at roughly 4.5× the per-email cost of Gemini 3.1 Flash-Lite. Use this when the upgraded precision materially reduces downstream review burden on a large corpus.
- **Enterprise governance:** Use Google Vertex AI / Gemini Enterprise Agent Platform or Microsoft Foundry deployment for stronger contractual data-handling terms, even though raw per-token pricing may be marginally higher than the direct provider APIs.

6. Quick Reference: Model Recommendations by Priority

Your Top Priority	Recommended Model	Provider	Notes
Highest F1 (any deployment)	Mistral Small 3.2 24B	Ollama	Test 1 F1 = 99.6% on 1,974 decided emails; 2.75 emails/s on Test 5; 24 GB VRAM
Highest F1 with near-complete decision rate	Llama 3.3 70B / Qwen 3.6 35B MoE	Ollama	Both reach F1 = 99.2%; Llama has 99.95% decision rate and needs 80+ GB VRAM for full speed; Qwen 35B MoE needs 32 GB VRAM
Highest Automation Yield (overall, tied)	Llama 3.3 70B / Grok 4.2 Non-Reasoning	Ollama / xAI API	AY = 99.85% for both; effectively no manual review queue
Highest Automation Yield (cloud)	Grok 4.2 Non-Reasoning	xAI API / Foundry	AY = 99.85%; F1 = 99.2%; ~$192 / 100K
Highest Automation Yield (lower-end offline)	Ministral 3 14B	Ollama	AY = 99.30%; fastest model in the benchmark; runs on 16 GB VRAM

Highest cloud F1	Grok 4.2 Non-Reasoning	xAI API	F1 = 99.2%; supplants Grok 4.1 Fast as the top cloud accuracy pick
Best practical accuracy (cloud)	Grok 4.2 Non-Reasoning	xAI API	F1 = 99.2%; AY = 99.85%; ~300,000 emails per weekend
Best value (cloud)	Gemini 3.1 Flash-Lite	Google AI Studio / Google Vertex AI / Gemini Enterprise Agent Platform	$0.25/1M input tokens via AI Studio; ~$42 per 100K emails; F1 = 96.0%; 100% on Tests 2, 3, and 4
Maximum speed (cloud)	Gemini 3.1 Flash-Lite (Agent Platform)	Google Vertex AI / Gemini Enterprise Agent Platform	1.72 emails/s on Test 5; ~384,000 emails/weekend; ~$43 / 100K
Maximum speed (offline)	Ministral 3 14B	Ollama	3.94 emails/s; ~879,000 emails/weekend; runs on 16 GB VRAM
Multi-category triage (cloud)	Gemini 3.1 Flash-Lite	Google AI Studio / Google Vertex AI / Gemini Enterprise Agent Platform	100% on Tests 2, 3, and 4; strongest low-cost cloud multilingual profile
Multi-category triage (offline)	Qwen 3.6 27B Dense	Ollama	100% on Tests 2, 3, and 4; accept low throughput (0.08 emails/s)
Most consistent across all four tests	Qwen 3.6 27B Dense	Ollama	1.2-point worst-to-best spread (narrowest of any retained model); 24 GB VRAM
Complete data privacy (balanced)	Gemma 4 26B Think	Ollama	F1 = 95.2%; 99.5% Test 2; 100%/99.3% Korean; 24+ GB VRAM; native 256K support with sufficient VRAM
Lower-end offline hardware	Ministral 3 14B	Ollama	F1 = 94.9% binary; fastest offline; 16+ GB VRAM; weak on multi-category
European data residency, including Switzerland or the UK	Claude 4.7 Opus	Amazon Bedrock (European regions)	Available in France, Germany, Ireland, Italy, Spain, Sweden, Switzerland, and the UK
EU multi-region, high-volume cloud	Gemini 3.1 Flash-Lite (Agent Platform)	Google Vertex AI / Gemini	eu deployment excludes UK and Switzerland; fastest cloud throughput; ~$43 / 100K

		Enterprise Agent Platform	
Korean multi-category (cloud)	Gemini 3.1 Flash-Lite / Gemini 3 Flash (preview) / Claude 4.7 Opus	Google / Bedrock	All hit 100% on Test 4; Gemini 3.1 Flash-Lite and Gemini 3 Flash (preview) also hit 100% on Test 3
Multilingual (incl. Korean, CJK)	Gemini 3.1 Flash-Lite (cloud) or Qwen 3.6 27B Dense / Gemma 4 26B Think (offline)	Google / Ollama	Flash-Lite and Qwen Dense are perfect across Tests 2–4; Gemma remains a strong fast-loading offline option
French-language focus	Mistral Small 3.2 24B or Ministral 3 14B	Ollama	French is a co-primary language in the Mistral family
Hybrid cost optimization	Gemini 3.1 Flash-Lite → Claude 4.7 Opus	Two-pass	Fast cull, then refine on Responsive + Inconclusive subset

7. Models to Avoid

Based on the May 2026 benchmark, the following models are explicitly excluded from the retained set and should not be used for forensic email responsiveness classification, despite remaining technically available:

- **OpenAI GPT-5.5**—refused to classify the benchmark emails, returning the error message *"This content was flagged for possible cybersecurity risk."* Unusable for forensic email work regardless of its underlying capability.
- **Grok 4.3**—the slowest cloud model in the benchmark by a wide margin (4 h 55 m on Test 1 versus 24 m for Grok 4.2 Non-Reasoning) and lower Test 1 F1 (98.33%) than its non-reasoning sibling. Dominated on every operational dimension by Grok 4.2 Non-Reasoning.
- **Mistral Large 3 (cloud)**—high abstention on binary responsiveness despite acceptable multi-category performance.
- **Qwen 2.5 (14B and 32B variants)**—collapsed precision on binary responsiveness; excluded for low accuracy.
- **Qwen 3.5 9B and Qwen 3.5 27B (Think and NoThink)**—superseded by Qwen 3.6.
- **Magistral 24B (offline)**—missed three true positives on Test 1; excluded for accuracy.
- **Nemotron 3 33B (offline)**—excluded for low accuracy and high false-positive rate on multi-category work.
- **GPT-OSS 20B (Low and High) and GPT-OSS 120B (Low and High)**—excluded either for low accuracy or because a smaller retained offline model (Gemma 4 26B Think) delivered equal or better accuracy at higher throughput and a far smaller VRAM footprint.
- **Gemma 4 31B (Think and NoThink)**—excluded because Gemma 4 26B Think delivered equal or better accuracy at higher throughput on the same hardware budget.
- **Gemma 4 E4B (Think and NoThink)**—excluded for low accuracy.

The following models are no longer part of the retained benchmark set because they have been superseded by newer or stronger siblings, but may still be available in Aid4Mail configurations:

- **Claude Opus 4.5, Claude Opus 4.6, Claude Sonnet 4.6, Claude Haiku 4.5**—superseded by Claude 4.7 Opus, which is the current Anthropic recommendation.
- **Gemini 2.5 Flash**—superseded by Gemini 3 Flash (preview) and 3.1 Flash-Lite.
- **OpenAI GPT-5.2**—superseded by GPT-5.4.
- **Grok 4.1 Fast and Grok 4.1 Fast+Reasoning**—superseded by Grok 4.2 Non-Reasoning; Grok 4.1 Fast is also slated for retirement on May 15, 2026.
- **Gemma 3 27B**—superseded by Gemma 4 26B Think on the same hardware budget.
- **Gemma 4 26B NoThink**—superseded by Gemma 4 26B (the reasoning/"Think" variant at 26B).

---

8. Before You Process: A Pre-Flight Checklist

Before running your first large-scale AI job, verify these items:

1. **Test on a small sample first.** Run 100–500 representative emails to verify that your model, prompt, and settings produce expected results.
2. **Check your prompt.** Use the **Verify** button in Aid4Mail's Project Settings to validate your prompt. Consider using a top commercial chat model (like Claude or GPT) to help refine your prompt before batch processing. The benchmark confirmed that prompt complexity is itself a variable in model behavior—not just a parameter to be optimized—so test across at least two models on a representative sample for sensitive matters. Test 5 and related Podesta-corpus testing showed that subjective themes—including absence-detection (inferring what an email deliberately is not saying) but not limited to it—produce wide variation across model families, with the same prompt yielding Responsive counts that differ by several-fold. Cross-model validation is essential whenever the responsiveness criterion involves judgment rather than concrete factual indicators.
3. **Enable incremental processing.** In Source settings, enable "Automatically record each email to allow incremental processing." If the job is interrupted, you can resume without reprocessing.
4. **Use pre-acquisition and post-acquisition filtering first.** Narrow your dataset with standard Aid4Mail filters before applying AI processing. This saves time and reduces costs.
5. **Consider attachment inclusion carefully.** The cost of including attachment text varies significantly with corpus type. On the 34,097-email Podesta Production Pilot, enabling document-text extraction increased input token volume by approximately 32% and added 7 minutes (Gemini 2.5 Flash) to 67 minutes (Mistral Small 3.2 24B) of processing time, while yielding only 11 additional Responsive classifications for Gemini and zero for Mistral. The benefit is corpus-specific. Note that Aid4Mail always includes each attachment's filename and extension in the payload, which often carries substantial signal on its own.
6. **For offline models:** Verify the server is running (`ollama serve` or LM Studio's Start Server), and confirm the context length matches in both the inference tool and Aid4Mail.

7. **For enterprise platforms:** Ensure your quotas are sufficient for the job size. Contact the platform provider if you need higher limits.
8. **Re-validate when models change.** Cloud model behavior shifts silently between revisions; offline models are reproducible and stable once downloaded. For long-running matters where mid-investigation behavioral drift would be unacceptable, prefer a fixed offline model or a pinned cloud model version with versioned API endpoints.

Date of publication: May 15, 2026.