
# Aid4Mail AI Classification Benchmark Report

*Report*

Fookes Software Ltd
Charmey, Switzerland
www.aid4mail.com

# Table of Contents

# Aid4Mail AI Classification Benchmark Report

*A practical evaluation of cloud and offline AI models for digital forensics and eDiscovery email classification.*

## 1. Overview and Purpose

This report presents the results of an independent, multi-test benchmark program comparing 11 AI models—5 commercial cloud models and 6 offline models—on realistic email-classification workloads relevant to digital forensics and eDiscovery practitioners.

The program included five numbered tests plus a separate large-scale Production Pilot. Tests 1–4 used synthetic or translated-synthetic responsive content for accuracy evaluation; real Podesta emails were used as the Test 1 unresponsive background and as the operational corpora for Test 5 and the Production Pilot. The tests measured classification accuracy (precision, recall, F1), multi-category discrimination capacity, multilingual performance on Korean content, and operational throughput at production scale.

The goal of this document is narrow and practical. It is intended to help Aid4Mail users:

1. Understand how each available model performs on realistic forensic and eDiscovery tasks.
2. Compare AI classification against traditional keyword search and Technology-Assisted Review (TAR).
3. Select the most appropriate model for a given workflow, given constraints on accuracy, cost, speed, data-residency, and available hardware.

Results are presented with the methodological caveats that apply to any classification benchmark. Nothing in this report should be read as a promise of a particular outcome in a specific matter; classifier performance in a given investigation depends on prevalence, the quality of the prompt, the nature of the corpus, and the model's fit to the task.

## 2. Quick Reference for Readers New to AI Classification

Before reading the results, these six metrics are worth knowing. They are used throughout the report.

- **Precision** measures the proportion of emails flagged Responsive that actually are responsive. High precision means few false positives (low over-collection).
- **Recall** measures the proportion of actually responsive emails the model found. High recall means few false negatives (low risk of missed evidence).

- **F1 score** is the harmonic mean of precision and recall. It penalizes extreme imbalances and is the standard combined effectiveness metric in TAR literature.
- **Decision rate** is the proportion of emails the model classified as Responsive or Unresponsive (i.e., not INCONCLUSIVE and not errored). A low decision rate indicates model abstention.
- **Automation Yield** is the proportion of the full corpus the model resolved correctly and confidently, computed as `(TP + TN) / Total`. Whereas F1 measures decision *quality* on items the model committed to, Automation Yield measures decision *coverage* across the corpus, including INCONCLUSIVE and errored items that still require downstream attention. A model that abstains frequently can post a high F1 while leaving meaningful work for human reviewers; Automation Yield captures that gap.
- **Prevalence** is the proportion of truly responsive emails in the full corpus. A 6% prevalence means that for every 100 emails reviewed, 6 are relevant. Prevalence has a strong effect on observed precision: at low prevalence, even small false-positive rates generate large false-positive counts in absolute terms.

The three-label system used in Aid4Mail—**Responsive**, **INCONCLUSIVE**, **Unresponsive**—lets the model flag genuinely ambiguous emails for human review rather than forcing a binary decision. INCONCLUSIVE results are counted as neither true positive nor true negative in the analysis; they represent the model's abstention boundary.

Precision, recall, and F1 throughout this report follow the textbook definitions—`TP / (TP + FP)` and `TP / (TP + FN)`—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 in the F1 calculation. This aligns the AI figures in this report with the same definitions used in the TREC Legal Track and TAR literature cited in Section 8.

## 3. Test Overview

Six tests were used for this benchmark. Tests 1 through 4 form the core accuracy benchmark. Production Pilot and Test 5 were both operational in nature: Production Pilot was an early large-scale pilot that validated weekend-throughput estimates and measured the impact of attachment inclusion, and Test 5 was used for throughput and cost analysis at production payload sizes, as the task itself proved difficult to specify reliably across all models. Both are covered in Section 7.

| Test | Corpus | Scope | Primary Purpose |
|---|---|---|---|
| **Production Pilot** | 34,097 Podesta emails | Three-label FOIA classification | Operational performance at realistic scale; attachment-inclusion impact |
| **Test 1** | 1,880 Podesta + 120 synthetic | Insider threat / data exfiltration | Realistic investigation scenario with 6% responsive prevalence |
| **Test 2** | 200 synthetic (40 per theme) | Five misconduct categories + clean/inconclusive | Multi-category discrimination across three languages |

| | | | |
|---|---|---|---|
| **Test 3** | 120 synthetic (Korean) | Insider threat / data exfiltration in Korean | Foreign-language classification capability |
| **Test 4** | 150 synthetic (Korean) | Five misconduct categories in Korean | Foreign-language multi-category discrimination |
| **Test 5** | 1,083 Podesta (March 2016) | Off-record-intent signal detection | Throughput and cost at production scale |

All tests were run on a single workstation with an AMD Ryzen 9 9950X3D, RTX 5090 (32 GB VRAM), and 192 GB DDR5. Offline models were served via Ollama using Q4_K_M quantization at 32K context length (except Production Pilot, noted where relevant). Cloud models were accessed through their respective provider APIs, with Anthropic's Claude routed via Amazon Bedrock.

Of the up to 40 models evaluated during the program, 11 were retained for the final benchmark set. The remainder were excluded because they underperformed, were superseded by a newer or stronger sibling, were unsuitable for the task, or offered no measurable advantage over a faster or cheaper alternative.

## Models Evaluated but Excluded from Results

The following models were tested during the program but are not included in the final results. Exclusion categories are informational; some models could fit more than one category.

**Superseded by a newer or better-performing sibling:**

- 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 (Ollama)
- Gemma 4 26B NoThink (Ollama)
- Grok 4.1 Fast
- Grok 4.1 Fast+Reasoning
- Grok 4.3

**Excluded based on benchmark performance (low accuracy, high abstention, or both):**

- Gemma 4 E4B Think (Ollama)
- Gemma 4 E4B NoThink (Ollama)
- GPT-OSS 20B High (Ollama)
- GPT-OSS 20B Low (Ollama)
- Magistral 24B (Ollama)
- Mistral Large 3
- Nemotron 3 33B (Ollama)
- OpenAI GPT-5.5
- Qwen 2.5 14B, JSON Schema variant (Ollama)
- Qwen 2.5 14B, Unstructured Output variant (Ollama)
- Qwen 2.5 32B (Ollama)

- Qwen 3.5 9B (Ollama)
- Qwen 3.5 27B Think (Ollama)
- Qwen 3.5 27B NoThink (Ollama)

**Excluded because a smaller offline model delivered equal or better accuracy at higher throughput and a far smaller VRAM footprint:**

- Gemma 4 31B Think (Ollama)
- Gemma 4 31B NoThink (Ollama)
- GPT-OSS 120B Low (Ollama)
- GPT-OSS 120B High (Ollama)—equal Test 1 F1 and Test 3 accuracy to Gemma 4 26B, lower Test 2 and Test 4 accuracy, roughly half the throughput, and requires ≥96 GB VRAM versus ≥24 GB for Gemma 4 26B.

Two specific exclusions are worth noting because they affected the cloud lineup directly:

- **Grok 4.3** was by far the slowest online model tested (4h 55m on Test 1 versus 24m for Grok 4.2 Non-Reasoning) and posted lower Test 1 accuracy than the non-reasoning variant of its predecessor. It is dominated on every operational dimension by Grok 4.2 Non-Reasoning.
- **OpenAI GPT-5.5** refused to classify the benchmark emails, returning the error message *"This content was flagged for possible cybersecurity risk."* The model is therefore unusable for this benchmark regardless of its underlying capability.

## 4. Summary Results

The table below consolidates headline accuracy results from Tests 1, 2, 3, and 4. Speed, throughput, and cost are covered separately in Sections 6 and 7. Models are ordered by Test 1 F1. Gemini 3.1 Flash-Lite is shown as two adjacent rows—one for the Google AI Studio deployment and one for the Gemini Enterprise Agent Platform deployment—because their AY, speed, and throughput differ even though Test 1 F1 and Tests 2/3/4 accuracy are identical.

| Model | Deployment | Test 1 F1 | Test 1 AY | Test 2 Acc. | Test 3 Acc. (KR) | Test 4 Acc. (KR) |
|---|---|---|---|---|---|---|
| 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% |
| 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% |

| | | | | | | |
|---|---|---|---|---|---|---|
| Claude 4.7 Opus (low effort) | Cloud | 97.2% | 99.65% | 100.0% | 99.2% | 100.0% |
| Gemini 3.1 Flash-Lite (Agent Platform) | Cloud | 96.0% | 99.50% | 100.0% | 100.0% | 100.0% |
| Gemini 3.1 Flash-Lite | Cloud | 96.0% | 99.45% | 100.0% | 100.0% | 100.0% |
| Gemma 4 26B | 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% |

For throughput, weekend processing volume, and cost at production scale, see Section 7.

**What stands out at a glance:**

- Every retained model in the benchmark achieved Test 1 F1 above 94%, placing all of them at or above the upper portion of the published Technology-Assisted Review (TAR 2.0 / CAL) range, and well above any keyword search baseline. See our Keyword Search TAR Performance Benchmarks report for context.
- The top five F1 scores span about 0.8 percentage points (99.6%, 99.2%, 99.2%, 99.2%, 98.8%) across two offline models, one cloud model, and a pair of new offline siblings (Qwen 3.6 27B Dense and 35B MoE), showing that no single deployment mode has a monopoly on accuracy. Three models—Llama 3.3 70B, Qwen 3.6 35B MoE, and Grok 4.2 Non-Reasoning—tie for second place at 99.17%.
- Recall is near the ceiling across the board: every retained model achieved 99% or higher recall on Test 1, meaning F1 differences between models are driven almost entirely by precision (over-collection), not by missed evidence.
- **Four entries scored a perfect 100.0% on Tests 2, 3, AND 4: Qwen 3.6 27B (Dense), Gemini 3 Flash, and both deployments of Gemini 3.1 Flash-Lite (AI Studio and Agent Platform).** Qwen 3.6 27B Dense is the only offline model in this group. Its Test 1 F1 (98.77%) and Automation Yield (99.80%) are also among the strongest, making it the most consistent performer across the four core tests.
- **F1 and Automation Yield tell different stories.** Mistral Small 3.2 24B sits at the top of an F1-ordered table (99.6% on decided emails) and OpenAI GPT-5.4 sits mid-table (97.6%), but both fall to the bottom on AY because of high INCONCLUSIVE counts (26 and 20 respectively). Conversely, Llama 3.3 70B and Grok 4.2 Non-Reasoning lead on AY (both at 99.85%) while tying for second on F1 alongside Qwen 3.6 35B MoE. The two metrics are complementary and should be read together; see §6.2 for the decision-coverage analysis.
- Throughput, weekend processing volumes, and production costs vary widely between models and are reported separately in Section 7 using Test 5 data where available, with Test 1 extrapolations for models not run on Test 5. Test 5 is more representative of real-world email payloads than Test 1.

# 5. Individual Model Analysis

This section summarizes each tested model's strengths, weaknesses, and the conditions under which it performs best. Models are grouped by deployment type (cloud vs. offline) and within each group ordered by Test 1 F1.

## 5.1 Cloud Models

### Grok 4.2 Non-Reasoning

- **Test 1:** Precision 98.3%, Recall 100.0%, F1 99.17%, Automation Yield 99.85%; one INCONCLUSIVE, zero errors.
- **Test 2:** 99.5% accuracy.
- **Test 3 (Korean):** 99.17% (one INCONCLUSIVE out of 120).
- **Test 4 (Korean):** 96.0%; struggled on compliance-violation classification (83.3% on that theme).
- **Speed:** 1.35 emails/s (Test 1); 24m 37s for 2,000 emails—third-fastest cloud model on Test 1, behind both Gemini 3.1 Flash-Lite deployments.
- **Cost:** $3.83 per 2,000 emails (~$192 per 100,000 Test 1 emails).

**Strengths:** Tied with Llama 3.3 70B for the highest Automation Yield in the benchmark (99.85%), tied for second-highest F1 in the benchmark (alongside Llama 3.3 70B and Qwen 3.6 35B MoE at 99.17%), and the highest-F1 cloud model in the retained set. Strong Korean handling on the focused binary task.

**Weaknesses:** Korean multi-category accuracy (96.0% on Test 4) is the weakest of the cloud models retained, with compliance-violation discrimination dragging the overall score. Substantially more expensive per email than Gemini 3.1 Flash-Lite at lower speed: pricing is $1.25/1M input and $2.50/1M output, versus $0.25/$1.50 for Flash-Lite (AI Studio) and $0.275/$1.65 for Flash-Lite (Agent Platform).

**Best fit:** Cloud workflows where high precision and high decision coverage both matter—insider-threat or exfiltration triage where missed evidence and over-collection are both costly. The natural successor to Grok 4.1 Fast for organizations that previously standardized on the Grok family, with materially better Korean multi-category accuracy at a higher per-email price.

### OpenAI GPT-5.4

- **Test 1:** Precision 95.2%, Recall 100.0%, F1 97.56%, Automation Yield 98.70%; 20 INCONCLUSIVE responses lowered the decision rate to 99.0%.
- **Test 2:** 100.0% accuracy.
- **Test 3 (Korean):** 98.33% (two INCONCLUSIVE).
- **Test 4 (Korean):** 98.0%; one weak category (Discrimination, 90.0%).
- **Speed:** 0.88 emails/s (Test 1), 37m 49s for 2,000 emails.
- **Cost:** $6.64 per 2,000 emails (~$332 per 100,000 Test 1 emails).

**Strengths:** Perfect recall on Test 1 and flawless multi-category accuracy on Test 2, with newly added Korean results that confirm strong multilingual performance (98.3% / 98.0%). Strong prompt sensitivity (responds well to detailed instructions).

**Weaknesses:** The highest INCONCLUSIVE count among retained cloud models (20 abstentions on Test 1 plus 2 on Test 3), though Mistral Small 3.2 24B has a higher Test 1 INCONCLUSIVE count overall. OpenAI GPT-5.4 is substantially more expensive than alternatives delivering similar or higher F1. At ~$383 per 100,000 emails on Test 5, it costs roughly 9× what Gemini 3.1 Flash-Lite costs for a closely matched F1; the cost–accuracy proposition is unfavorable for routine classification.

**Best fit:** Organizations already standardized on the OpenAI ecosystem, or analysis tasks where the model's strong summarization and reasoning capabilities justify the price. Not the best choice for cost-sensitive classification.

### Claude 4.7 Opus (low effort)

- **Test 1:** Precision 94.5%, Recall 100.0%, F1 97.17%, Automation Yield 99.65%; zero INCONCLUSIVE, zero errors.
- **Test 2:** 100.0% accuracy.
- **Test 3 (Korean):** 99.17% (one error out of 120).
- **Test 4 (Korean):** 100.0%—perfect.
- **Speed:** 0.47 emails/s (Test 1), 1h 11m 20s for 2,000 emails.
- **Cost:** $25.50 per 2,000 emails (~$1,275 per 100,000 Test 1 emails)—by far the highest cost in the set.

**Strengths:** Perfect recall on Test 1, perfect Test 2 accuracy, and a perfect 100.0% on the Korean multi-category task—the most demanding multilingual test in the benchmark. Decided every email on Test 1 with zero INCONCLUSIVE, the cleanest decision-coverage profile of any retained cloud model. Most capable model in the set for reasoning-heavy work outside the scope of classification alone (e.g., summarization, long-form analysis, Korean translation—used extensively during this benchmark's corpus preparation).

**Weaknesses:** Substantially slower than other cloud alternatives and roughly 28× the cost of Gemini 3.1 Flash-Lite for similar F1. Even at the "low effort" setting it consumes considerably more input tokens than Claude Opus 4.6 did on the same task (4.92M vs. 2.80M), pushing per-email cost up further.

**Best fit:** Analysis tasks that benefit from Opus' broader capabilities (email summarization, multi-document reasoning, language translation), or high-value matters where cost is not a constraint and the reviewer wants the best available reasoning model behind every decision. Not the right choice for large-volume routine classification.

### Gemini 3.1 Flash-Lite

Gemini 3.1 Flash-Lite is now generally available (non-preview) and was tested in two deployments: the Google AI Studio direct API and the Gemini Enterprise Agent Platform (EU region). The two deployments produce identical Test 1 F1 (96.0%) and perfect scores on Tests 2, 3, and 4. They differ in speed, weekend throughput, per-token pricing, and Automation Yield (by 0.05 pts on Test 1). The figures below split by deployment where the difference is material; accuracy claims apply to both.

- **Test 1 (both deployments):** Precision 92.3%, Recall 100.0%, F1 96.00%, 10 false positives.
  - **AI Studio:** Automation Yield 99.45%; one INCONCLUSIVE, zero errors.
  - **Agent Platform:** Automation Yield 99.50%; zero INCONCLUSIVE, zero errors, 100% decision rate.
- **Test 2 (both deployments):** 100.0% accuracy, perfect across all five misconduct categories.
- **Test 3 (Korean, both deployments):** 100.0%—perfect.
- **Test 4 (Korean, both deployments):** 100.0%—perfect.
- **Speed:**
  - **AI Studio:** 1.40 emails/s (Test 1), 23m 47s for 2,000 emails; 1.30 emails/s on Test 5.
  - **Agent Platform:** 1.72 emails/s (Test 1), 19m 24s for 2,000 emails—the fastest cloud model on Test 1; 1.72 emails/s on Test 5, finishing the Test 5 corpus about 25% sooner than AI Studio (or about 33% higher emails/s).
- **Cost:**
  - **AI Studio:** $0.89 per 2,000 emails (~$45 per 100,000 Test 1 emails; ~$42 per 100,000 Test 5 emails). Pricing $0.25/1M input, $1.50/1M output.
  - **Agent Platform:** $0.90 per 2,000 emails (~$45 per 100,000 Test 1 emails; ~$43 per 100,000 Test 5 emails). Pricing $0.275/1M input, $1.65/1M output.

**Strengths:** Perfect 100.0% on Tests 2, 3, AND 4—shared in the retained set only with Qwen 3.6 27B Dense (offline) and Gemini 3 Flash (cloud). Very strong multilingual performance, competitive speed, and the lowest cloud cost in the benchmark. The most consistently high-performing low-cost cloud model across the benchmark set. The Agent Platform deployment is the fastest cloud model on Test 1 and on Test 5.

**Weaknesses:** Precision is below Grok 4.2 Non-Reasoning, OpenAI GPT-5.4, and Claude 4.7 Opus on Test 1 because of 10 false positives at 100% recall. Available on Google Vertex AI / Gemini Enterprise Agent Platform only in three deployments (`global`, `us`, and `eu`); Aid4Mail surfaces the `eu` and `us` deployments. The `eu` deployment is a multi-region endpoint that does not include the UK or Switzerland. Customers requiring UK or Swiss data residency for Gemini 3.1 Flash-Lite have no enterprise option at this time.

**Best fit:** The default recommendation for cloud-based Aid4Mail workflows, particularly when the corpus may contain multilingual content or the task involves multi-category triage. Use the Agent Platform deployment when the project is eligible for the US or EU multi-region endpoint and speed matters.

### Gemini 3 Flash (preview)

- **Test 1:** Precision 89.6%, Recall 100.0%, F1 94.49%, Automation Yield 99.25%.
- **Test 2:** 100.0% accuracy.
- **Test 3 (Korean):** 100.0%.
- **Test 4 (Korean):** 100.0%—perfect.
- **Speed:** 1.09 emails/s (Test 1), 30m 33s for 2,000 emails.
- **Cost:** $1.62 per 2,000 emails (~$81 per 100,000 Test 1 emails; ~$119 per 100,000 Test 5 emails).

**Strengths:** Perfect scores on Tests 2, 3, AND 4—shared in the retained set with Qwen 3.6 27B Dense and both deployments of Gemini 3.1 Flash-Lite. Reasonable throughput. A

larger-model alternative to Flash-Lite when a workflow benefits from that model profile, though this benchmark did not show an accuracy or context-window advantage over Flash-Lite on the retained tests.

**Weaknesses:** The lowest Test 1 F1 of the retained cloud models—14 false positives versus 10 for Flash-Lite—and higher cost than Flash-Lite: about 1.8× on Test 1 and about 2.8× on Test 5. Its Test 5 speed is close to the Flash-Lite AI Studio deployment but materially below the Agent Platform deployment. Still in preview.

**Best fit:** Cloud workflows that benefit from Gemini 3 Flash's larger-model profile, particularly where an organization wants an alternative to Flash-Lite for multi-category or multilingual triage. Not a precision upgrade over Flash-Lite on adversarial binary classification, and this benchmark did not show a context-window advantage over Flash-Lite.

## 5.2 Offline Models

Offline models carry no per-email cost, but they do require local GPU hardware. The reference workstation for this benchmark cost CHF 4,967 (November 2025) with its RTX 5090 and 192 GB of DDR5. Organizations weighing offline deployment should factor in hardware amortization and electricity, which are not reflected in the "$0.00" cost column.

### Mistral Small 3.2 24B (Ollama)

- **Test 1:** Precision 99.2%, Recall 100.0%, F1 99.58% over the 1,974 decided emails—the highest F1 in the benchmark; Automation Yield 98.65% over the full 2,000, reflecting 26 INCONCLUSIVE responses at a 98.7% decision rate.
- **Test 2:** 97.5% accuracy.
- **Test 3 (Korean):** 97.5%.
- **Test 4 (Korean):** 97.33%.
- **Speed:** 2.60 emails/s (Test 1)—the fastest offline model in its accuracy tier. 12m 50s for 2,000 emails.

**Strengths:** Highest Test 1 F1 in the benchmark on decided emails, excellent throughput for an offline model, consistent multilingual accuracy, and a reasonable VRAM footprint (24 GB). Well matched to workflows that treat the three-label design as a safety feature.

**Weaknesses:** Produces more abstentions than any other retained model at 32K context—26 INCONCLUSIVE responses on Test 1 versus at most 1 for the other offline models—so the headline F1 is computed over 1,974 decided emails rather than the full 2,000. Workflows that cannot accommodate a small human-review queue of flagged-for-decision items will prefer a model with a higher decision rate.

**Best fit:** Under textbook metrics the top-F1 model in the benchmark on decided emails. Well suited to high-volume binary classification on 24 GB or 32 GB GPUs where the three-label abstention boundary is treated as a feature and reviewers can absorb a modest INCONCLUSIVE queue. Less well suited to workflows requiring near-100% decision rates without human review.

### Llama 3.3 70B (Ollama)

- **Test 1:** Precision 98.4%, Recall 100.0%, F1 99.17%—tied for second place behind Mistral Small 3.2 24B alongside Qwen 3.6 35B MoE and Grok 4.2 Non-Reasoning; Automation Yield 99.85%—tied with Grok 4.2 Non-Reasoning for the highest in the benchmark.
- **Test 2:** 98.5% accuracy.
- **Test 3 (Korean):** 95.83%; four Unresponsive classifications and one INCONCLUSIVE, the highest count of non-Responsive outcomes among retained models in that test.
- **Test 4 (Korean):** 93.33%—second-weakest Korean multi-category score in the retained set, behind Ministral 3 14B at 91.33%; corruption & bribery is Llama's weakest Test 4 category at 80.0%.
- **Speed:** 0.14 emails/s (Test 1), 3h 55m for 2,000 emails on the 32 GB reference workstation. This figure reflects partial CPU offload, which the model requires at reasonable context lengths on 32 GB of VRAM. On an 80 GB-class GPU (for example, a single NVIDIA H100) the model can be kept fully GPU-resident, and throughput is expected to increase substantially. We have not benchmarked this configuration directly.

**Strengths:** Tied for second-highest F1 in the benchmark (alongside Qwen 3.6 35B MoE and Grok 4.2 Non-Reasoning) and near-perfect English-language precision at 100% recall. Strong multi-category accuracy. Decided on every responsive email in Test 1 with only one INCONCLUSIVE routing, which is operationally distinct from Mistral Small's headline F1 (99.6% over 1,974 decided emails versus Llama's 99.2% over 1,999 decided emails, with only one INCONCLUSIVE routing).

**Weaknesses:** Very slow on the reference hardware because the 70B model does not fit comfortably in 32 GB of VRAM at reasonable context lengths, forcing partial CPU offload. Korean handling shows measurable degradation relative to smaller offline models such as Gemma 4 26B and the Qwen 3.6 family. Requires substantially more VRAM (≥80 GB recommended) to run efficiently. Practitioners planning to deploy this model should size their GPU accordingly; hybrid CPU/GPU inference is viable for scheduled overnight runs but not for production-rate throughput.

**Best fit:** Air-gapped or data-residency-constrained environments where English-language classification accuracy is paramount and throughput can be scheduled around overnight or weekend runs on appropriate hardware. Not recommended on 32 GB GPUs for production throughput, and not the first choice when the corpus has substantial Korean content.

### Qwen 3.6 27B Dense (Ollama)

- **Test 1:** Precision 97.6%, Recall 100.0%, F1 98.77%, Automation Yield 99.80%; one INCONCLUSIVE, zero errors.
- **Test 2:** 100.0% accuracy—perfect.
- **Test 3 (Korean):** 100.0%—perfect.
- **Test 4 (Korean):** 100.0%—perfect.
- **Speed:** 0.08 emails/s (Test 1), 6h 38m for 2,000 emails—the slowest retained model on the reference workstation.

**Strengths:** The only offline model in the perfect-on-Tests-2/3/4 group (the other three entries are cloud models: Gemini 3 Flash (preview) and both deployments of Gemini 3.1 Flash-Lite). Combined with a strong Test 1 F1 (98.77%) and AY of 99.80%, this is the most

consistent performer in the entire retained set—best-to-worst spread of just 1.2 points across the four core tests, narrowly ahead of Qwen 3.6 35B MoE at 1.5 points. Exceptional multilingual and multi-category capability. Supports a 256K native context window.

**Weaknesses:** The slowest retained model on the reference hardware. Throughput is impractical for production-scale workloads; a 100,000-email run would take roughly two weeks of continuous processing on the reference workstation. Limited operational suitability without faster hardware.

**Best fit:** Small, high-stakes corpora where accuracy matters more than throughput, particularly multilingual matters with Korean, French, Spanish, or German content. Strong choice for air-gapped environments where overnight or multi-day runs are acceptable on small-to-medium corpora.

### Qwen 3.6 35B MoE (Ollama)

- **Test 1:** Precision 98.4%, Recall 100.0%, F1 99.17%, Automation Yield 99.80%; one INCONCLUSIVE, one error.
- **Test 2:** 99.5% accuracy.
- **Test 3 (Korean):** 99.17%.
- **Test 4 (Korean):** 98.0%.
- **Speed:** 0.16 emails/s (Test 1), 3h 35m for 2,000 emails—roughly 2× faster than the Qwen 3.6 27B Dense variant thanks to the MoE architecture (35B parameters with ~3B active per token).

**Strengths:** Tied for second-highest F1 in the benchmark (99.17%) alongside Llama 3.3 70B and Grok 4.2 Non-Reasoning, with the same Automation Yield (99.80%) as the Dense sibling. The second most consistent retained model on the four-test spread (1.5 points between best and worst), narrowly behind Qwen 3.6 27B Dense at 1.2 points. Roughly twice the throughput of the Dense variant for a small accuracy concession on Tests 2, 3, and 4. Strong Korean performance. Supports a 256K native context window.

**Weaknesses:** Slower than Mistral Small 3.2 24B and most cloud alternatives. The MoE architecture has a large parameter footprint despite low active-parameter count, requiring more VRAM than the Dense variant.

**Best fit:** Air-gapped multilingual workflows where accuracy and throughput both matter. The right pick when the Dense 27B is too slow for the corpus size but the workflow still demands the Qwen 3.6 family's multilingual and multi-category accuracy.

### Gemma 4 26B (Ollama)

- **Test 1:** Precision 90.9%, Recall 100.0%, F1 95.24%, Automation Yield 99.30%.
- **Test 2:** 99.5% accuracy.
- **Test 3 (Korean):** 100.0%.
- **Test 4 (Korean):** 99.33%—the best Korean multi-category result of any model tested outside the perfect-scoring set (Gemini 3 Flash, Gemini 3.1 Flash-Lite in both deployments, Claude 4.7 Opus, Qwen 3.6 27B Dense).
- **Speed:** 0.14 emails/s (Test 1), 3h 53m for 2,000 emails.

**Strengths:** Well-balanced offline model with near-perfect accuracy across Tests 2, 3, and 4, including the strongest Korean multi-category performance among the previously retained offline models. Offers cloud-class accuracy for air-gapped environments, in the same 24 GB

VRAM class as Qwen 3.6 27B Dense and smaller than Qwen 3.6 35B MoE. Supports a 256K native context window.

**Weaknesses:** Second-lowest Test 1 precision in the offline tier—12 false positives, versus 1 for Mistral Small, 2-3 for the Qwen 3.6 models, and 13 for Ministral 3 14B (the only retained offline model with lower precision). Slow on the reference hardware. Now superseded on Tests 2/3/4 by Qwen 3.6 27B Dense, which posts perfect scores on all three tests.

**Best fit:** A solid offline option for organizations whose hardware favors a 24 GB-class model and who want strong multilingual performance without committing to the slower Qwen 3.6 27B Dense or the larger 35B MoE. Particularly well suited to mid-size air-gapped deployments with mixed-language email content.

### Ministral 3 14B (Ollama)

- **Test 1:** Precision 90.2%, Recall 100.0%, F1 94.86%, Automation Yield 99.30%.
- **Test 2:** 93.0% accuracy—the lowest of the retained models.
- **Test 3 (Korean):** 98.33%; the fastest retained model above 98%.
- **Test 4 (Korean):** 91.33%; struggled on compliance-violation classification (73.33%).
- **Speed:** 3.80 emails/s (Test 1)—the fastest model in the benchmark, cloud or offline. 8m 46s for 2,000 emails.

**Strengths:** Remarkable throughput, native 256K context-length support, and a small enough memory footprint (16 GB VRAM) to run on entry-level GPUs. Competitive F1 on the focused binary task in both English and Korean.

**Weaknesses:** Accuracy drops noticeably on multi-category discrimination tasks, consistent with its smaller parameter count. Tends to over-flag on compliance-violation themes where semantic overlap with other categories is high.

**Best fit:** High-throughput triage on binary classification tasks, on-premises deployments with modest GPU budgets, or any scenario where the classification task can be decomposed into focused binary passes rather than a single wide multi-category call.

---

# 6. Comparative Analysis

## 6.1 Accuracy and Consistency

No single model dominated across all four accuracy tests, but **Qwen 3.6 27B Dense** comes the closest: it posted perfect scores on Tests 2, 3, and 4 and an F1 of 98.77% on Test 1. Mistral Small 3.2 24B earned the highest single F1 (99.58% on Test 1) over its 1,974 decided emails. Three models—Llama 3.3 70B, Qwen 3.6 35B MoE, and Grok 4.2 Non-Reasoning—tie for second at 99.17% with near-complete decision coverage. Grok 4.2 Non-Reasoning is the highest-F1 cloud model in the retained set (99.17%) and ties Llama 3.3 70B for the highest Automation Yield (99.85%).

A useful way to think about consistency is to look at the spread between each model's best and worst test result. Smaller is better:

| Model | Best Test Result | Worst Test Result | Spread |
|---|---|---|---|
| Qwen 3.6 27B (Dense) | 100.0% (Tests 2, 3, 4) | 98.8% (Test 1 F1) | 1.2 pts |
| Qwen 3.6 35B (MoE) | 99.5% (Test 2) | 98.0% (Test 4) | 1.5 pts |
| Mistral Small 3.2 24B | 99.6% (Test 1 F1) | 97.3% (Test 4) | 2.3 pts |
| OpenAI GPT-5.4 | 100.0% (Test 2) | 97.6% (Test 1 F1) | 2.4 pts |
| Claude 4.7 Opus (low effort) | 100.0% (Tests 2, 4) | 97.2% (Test 1 F1) | 2.8 pts |
| Grok 4.2 Non-Reasoning | 99.5% (Test 2) | 96.0% (Test 4) | 3.5 pts |
| Gemini 3.1 Flash-Lite (Agent Platform) | 100.0% (Tests 2, 3, 4) | 96.0% (Test 1 F1) | 4.0 pts |
| Gemini 3.1 Flash-Lite | 100.0% (Tests 2, 3, 4) | 96.0% (Test 1 F1) | 4.0 pts |
| Gemma 4 26B | 100.0% (Test 3) | 95.2% (Test 1 F1) | 4.8 pts |
| Gemini 3 Flash (preview) | 100.0% (Tests 2, 3, 4) | 94.5% (Test 1 F1) | 5.5 pts |
| Llama 3.3 70B | 99.2% (Test 1 F1) | 93.3% (Test 4) | 5.9 pts |
| Ministral 3 14B | 98.3% (Test 3) | 91.3% (Test 4) | 7.0 pts |

The two new Qwen 3.6 siblings sit at the top of the consistency table. Qwen 3.6 27B Dense in particular is the only model that combined a perfect or near-perfect score on every test—a meaningful signal of reliability and predictability across tasks and languages.

A second observation is that all retained models in the set achieved 99%+ recall on Test 1. The differentiator between F1 scores is almost entirely precision—how much over-collection the model introduces. Models that achieved top F1 did so by returning fewer false positives, not by finding more true positives.

## 6.2 Decision Coverage

F1 measures decision quality on items the model committed to; it is silent on items the model deferred via INCONCLUSIVE or failed outright. Automation Yield complements F1 by measuring the proportion of the full Test 1 corpus that the model resolved correctly and confidently.

| Model | Deployment | Test 1 F1 | Test 1 AY |
|---|---|---|---|
| Llama 3.3 70B | Offline | 99.17% | 99.85% |
| Grok 4.2 Non-Reasoning | Cloud | 99.17% | 99.85% |
| Qwen 3.6 27B (Dense) | Offline | 98.77% | 99.80% |
| Qwen 3.6 35B (MoE) | Offline | 99.17% | 99.80% |
| Claude 4.7 Opus (low effort) | Cloud | 97.17% | 99.65% |
| Gemini 3.1 Flash-Lite (Agent Platform) | Cloud | 96.00% | 99.50% |
| Gemini 3.1 Flash-Lite | Cloud | 96.00% | 99.45% |

| | | | |
|---|---|---|---|
| Gemma 4 26B | Offline | 95.24% | 99.30% |
| Ministral 3 14B | Offline | 94.86% | 99.30% |
| Gemini 3 Flash (preview) | Cloud | 94.49% | 99.25% |
| OpenAI GPT-5.4 | Cloud | 97.56% | 98.70% |
| Mistral Small 3.2 24B | Offline | 99.58% | 98.65% |

Three patterns stand out. First, **Llama 3.3 70B and Grok 4.2 Non-Reasoning are tied for the top yield at 99.85%**—1,997 of 2,000 emails fully resolved—both also tying for second on F1 alongside Qwen 3.6 35B MoE. Both committed to a verdict on essentially every email with at most one INCONCLUSIVE routing. Second, **Claude 4.7 Opus is the second-best cloud model on yield at 99.65%**, posting a 100% decision rate on Test 1—every verdict was either correct or a committed false positive. Third, **the Qwen 3.6 family (both Dense and MoE) sits at 99.80%**, pairing top-tier F1 with near-complete decision coverage.

F1 and Automation Yield should be read together. A model with high F1 and high yield decides broadly and accurately. A model with high F1 but lower yield (Mistral Small 3.2 24B, OpenAI GPT-5.4) decides selectively but accurately—appropriate where the three-label system is treated as a safety feature and reviewers absorb the INCONCLUSIVE queue. A model with high yield but lower F1 decides broadly but commits more false positives—faster through the pipeline, more manual rejection at review.

## 6.3 Speed

Test 5 processing times for the 1,083-email corpus are the best reference for realistic throughput expectations: the corpus has no upper size filter on individual emails, and the average message size of 81 KB is likely above the typical personal or business inbox average. Test 1 emails were filtered to 1–68 KB to keep total benchmark runtime manageable across all evaluated models. For models measured in both tests, Test 5 emails/s was equal to or higher than Test 1 in seven of eight deployment-rows, so Test 1 throughput should not be treated as a universal upper bound on production behavior.

| Model | Test 5 duration | Test 5 emails/s |
|---|---|---|
| Ministral 3 14B | 4m 35s | 3.94 |
| Mistral Small 3.2 24B | 6m 34s | 2.75 |
| Gemini 3.1 Flash-Lite (Agent Platform) | 10m 29s | 1.72 |
| Gemini 3.1 Flash-Lite | 13m 55s | 1.30 |
| Gemini 3 Flash (preview) | 14m 47s | 1.22 |
| OpenAI GPT-5.4 | 18m 21s | 0.98 |
| Gemma 4 26B | 1h 28m 42s | 0.20 |
| Llama 3.3 70B | 2h 8m 42s | 0.14 |

Test 5 was conducted before Claude 4.7 Opus, Grok 4.2 Non-Reasoning, Qwen 3.6 27B Dense, and Qwen 3.6 35B MoE were added to the benchmark. Their throughput figures

below are Test 1-only estimates; production-payload throughput has not been directly measured for these models.

| Model | Test 1 duration | Test 1 emails/s |
|---|---|---|
| Grok 4.2 Non-Reasoning | 24m 37s | 1.35 |
| Claude 4.7 Opus (low effort) | 1h 11m 20s | 0.47 |
| Qwen 3.6 35B (MoE) | 3h 34m 51s | 0.16 |
| Qwen 3.6 27B (Dense) | 6h 37m 52s | 0.08 |

Two patterns emerge. First, the fastest offline model (Ministral 3 14B) runs ahead of every retained cloud model on the reference workstation, because small offline models operate at local network latency rather than cloud round-trip latency. Second, offline throughput scales with payload size in a way that cloud models do not—cloud throughput is largely bound by API overhead and is comparatively stable regardless of email size, while local inference speed depends directly on the actual token count. See Section 7 for weekend-throughput projections based on these figures.

## 6.4 Cost

Cloud pricing per million tokens (at the time of testing) and cost per 100,000 emails based on Test 1 token volumes:

| Model | Input $/1M | Output $/1M | Cost / 100K emails (Test 1) |
|---|---|---|---|
| Gemini 3.1 Flash-Lite | $0.25 | $1.50 | ~$45 |
| Gemini 3.1 Flash-Lite (Agent Platform) | $0.275 | $1.65 | ~$45 |
| Gemini 3 Flash (preview) | $0.50 | $3.00 | ~$81 |
| Grok 4.2 Non-Reasoning | $1.25 | $2.50 | ~$192 |
| OpenAI GPT-5.4 | $2.50 | $15.00 | ~$332 |
| Claude 4.7 Opus (low effort) | $5.00 | $25.00 | ~$1,275 |

Test 1 emails were filtered to 1–68 KB; production payloads (Test 5) average around 81 KB. Per-email token volumes between Test 1 and Test 5 are model-dependent: approximately +2% to +4% for retained offline models, −5% to −6% for Gemini 3.1 Flash-Lite in both deployments, +16% for OpenAI GPT-5.4, and +50% for Gemini 3 Flash. The Test 5 cost projections in Section 7.2 reflect these per-model deltas.

At 100,000-email scale, the difference between the cheapest and most expensive cloud model is roughly $1,230—the difference between a trivial line item and a meaningful budget expense. For offline models, the marginal cost per email is zero once the hardware is in place; a workstation in the $5,000 range amortizes quickly against Claude 4.7 Opus pricing and very slowly against Gemini 3.1 Flash-Lite pricing.

Cost-per-email is not by itself a measure of cost-effectiveness. A model that produces more false positives per dollar spent still costs more in reviewer hours downstream. Combining F1 with cost produces a more useful picture.

## 6.5 Cost-effectiveness (Accuracy per Dollar)

The table below pairs Test 1 F1 with Test 1 cost per 100,000 emails, ordered from cheapest to most expensive. Offline models are excluded because their per-email cost is zero, making them uniformly dominant on this metric at the margin.

| Model | F1 | Cost / 100K emails (Test 1) | Notes |
|---|---|---|---|
| Gemini 3.1 Flash-Lite | 96.0% | ~$45 | Perfect on Tests 2–4 at the lowest cloud price point; available in two deployments (AI Studio and Agent Platform) at nearly identical cost |
| Gemini 3 Flash (preview) | 94.5% | ~$81 | Larger-model alternative to Flash-Lite; identical perfect scores on Tests 2, 3, 4 but lower Test 1 F1; ~1.9× the cost of Flash-Lite. |
| Grok 4.2 Non-Reasoning | 99.2% | ~$192 | Highest cloud F1 and tied-best AY—best accuracy per dollar above the $100/100K threshold |
| OpenAI GPT-5.4 | 97.6% | ~$332 | F1 close to Grok 4.2 Non-Reasoning at ~1.7× the cost |
| Claude 4.7 Opus (low effort) | 97.2% | ~$1,275 | Premium pricing not justified for straightforward classification |

Gemini 3.1 Flash-Lite is the clear value leader in the cloud tier, with Grok 4.2 Non-Reasoning pulling ahead on raw accuracy and decision coverage at roughly 4× the cost. Claude 4.7 Opus and OpenAI GPT-5.4 deliver strong accuracy but at a cost that is only justified when the model is serving capabilities beyond classification—Claude in particular excels at analysis and translation tasks that were used extensively in preparing this benchmark.

## 6.6 Multilingual Performance

Tests 3 and 4 probed Korean-language handling, a challenging case for most AI models because Korean is typologically distant from English and uses a non-Latin script. Results:

- On the focused binary task (Test 3), five model entries achieved 100%: both deployments of Gemini 3.1 Flash-Lite (AI Studio and Agent Platform), Gemini 3 Flash, Gemma 4 26B, and Qwen 3.6 27B Dense.
- On the multi-category task (Test 4), five model entries achieved 100%: both deployments of Gemini 3.1 Flash-Lite, Gemini 3 Flash, Claude 4.7 Opus, and Qwen 3.6 27B Dense. Gemma 4 26B was close behind at 99.33%—the best Korean multi-category result of any model outside the perfect-scoring set.
- **Qwen 3.6 27B Dense, Gemini 3 Flash, and both deployments of Gemini 3.1 Flash-Lite all scored 100% on Tests 2, 3, AND 4**—a clean sweep of the multi-category and multilingual benchmarks. Qwen 3.6 27B Dense is the only offline model in this group.
- Llama 3.3 70B showed the weakest Test 3 result in the retained set (95.83%) and the second-weakest Test 4 result (93.33%, behind Ministral 3 14B at 91.33%) despite a tied-second English F1, a reminder that English-language benchmarks do not always predict multilingual behavior.

- Grok 4.2 Non-Reasoning was the weakest cloud model on Test 4 (96.0%), with compliance violations the dragging category at 83.3%.

For French and Spanish, Test 2 included five emails per theme in each language (25 per language overall). All retained models maintained at least 93% accuracy on this multilingual subset, though the small sample size limits the granularity of that conclusion.

---

# 7. Production-Scale Operational Tests

Two tests in this benchmark were designed to measure operational performance rather than classification accuracy: Production Pilot, an early large-scale pilot on 34,097 Podesta emails, and Test 5, conducted on 1,083 real Podesta emails with no upper size filter. Production Pilot is the only test in the program at full production scale and the only one that measured the impact of attachment inclusion; Test 5 provides the reference throughput and cost figures cited throughout this report.

## 7.1 Production Pilot: Large-Scale FOIA Classification

Production Pilot was conducted on March 2, 2026, before Gemini 3.1 Flash-Lite was released. It used Gemini 2.5 Flash (since superseded by the more accurate Gemini 3.1 Flash-Lite) and Mistral Small 3.2 24B. A future run with the newer Gemini model is possible, but the results below remain useful for their operational insights.

A pre-filtered subset of the publicly available John Podesta Emails corpus—34,097 personal, non-duplicate emails (3.62 GB), reduced from 50,887 originals using the Aid4Mail filter query `Type:Personal AND NOT Type:Duplicate`—was classified into three FOIA-style categories: Responsive, Unresponsive, and INCONCLUSIVE. The corpus consists primarily of English-language political campaign correspondence leaked and published by WikiLeaks in 2016, and is widely used as a reference dataset by forensics and eDiscovery practitioners. No pre-verified ground truth was established, so Production Pilot results cannot be used to evaluate classification accuracy. The purpose was to measure operational performance at realistic production scale and to assess the impact of attachment inclusion on classification outcomes.

Each model processed the corpus twice: once excluding attachment data, and once including extracted document text. Gemini 2.5 Flash was accessed via Google Vertex AI (europe-west1, Belgium). Mistral Small 3.2 24B was run locally via Ollama on the reference workstation, with context length set to 32K when excluding attachments and 64K when including document contents.

The classification prompt applied a detailed FOIA-style political campaign review, instructing models to mark an email as Responsive only when participants—not quoted or forwarded third parties—explicitly discussed, authorized, or acknowledged conduct involving deliberate deception, circumvention of legal or regulatory requirements, improper coordination with nominally independent entities, unauthorized disclosure of confidential information, or concealment of recognized wrongdoing. The prompt specified that evidence must be direct and unambiguous, not merely inferable from political context or standard campaign practices, and directed models to default to Unresponsive when in doubt.

**Results by run:**

| Metric | Gemini 2.5 Flash (no attachments) | Gemini 2.5 Flash (with documents) | Mistral Small 3.2 24B (no attachments) | Mistral Small 3.2 24B (with documents) |
|---|---|---|---|---|
| Context length | 1,024K | 1,024K | 32K | 64K |
| Duration | 5h 12m 21s | 5h 19m 21s | 4h 34m 16s | 5h 41m 23s |
| Responsive | 136 | 147 | 11 | 11 |
| INCONCLUSIVE | 207 | 208 | 278 | 267 |
| Unresponsive | 33,754 | 33,742 | 33,808 | 33,819 |
| Input tokens | 50,910,551 | 67,127,060 | 49,129,602 | 65,201,450 |
| Output tokens | 204,645 | 204,729 | 409,135 | 408,654 |
| Speed (emails/s) | 1.82 | 1.78 | 2.07 | 1.66 |
| Speed (tokens/s) | 2,727 | 3,514 | 3,010 | 3,203 |
| Cost | $15.78 | $20.65 | $0.00 | $0.00 |

**Combined results (both runs):**

| Metric | Gemini 2.5 Flash | Mistral Small 3.2 24B |
|---|---|---|
| Total duration | 10h 31m 42s | 10h 15m 39s |
| Total input tokens | 118,037,611 | 114,331,052 |
| Total output tokens | 409,374 | 817,789 |
| Speed (emails/s) | 1.80 | 1.85 |
| Speed (tokens/s) | 3,125 | 3,117 |
| Average email context length | 1,731 | 1,677 |
| Total cost | $36.43 | $0.00 |
| Weekend throughput (est.) | ~400,000 | ~410,000 |

“Weekend throughput” estimates the number of emails classifiable during a 62-hour unattended run (Friday 18:00 to Monday 08:00). Actual performance varies with hardware specifications, AI platform region, payload size and nature, and prompt complexity.

**Observations from Production Pilot:**

- At full production scale, both models sustained throughput close to 2 emails/s, with weekend estimates of roughly 400,000 emails. These real-world figures are broadly consistent with the extrapolated weekend throughput reported in Section 7.2 below.
- Including extracted document text from attachments increased input-token volume by roughly 32% for both models and modestly changed the Responsive/INCONCLUSIVE distribution. Gemini 2.5 Flash flagged 11 additional emails as Responsive when attachments were included; Mistral Small 3.2 24B produced 11 fewer INCONCLUSIVE responses. For matters where attachment

contents are likely to change the classification decision, this trade-off is worth measuring on a representative sample before committing to a full run.

- Production-scale cost for the cheapest cloud path (Gemini 2.5 Flash, no attachments) was $15.78 to classify 34,097 emails—roughly $46 per 100,000 emails. This is the same order of magnitude as the Test 5 projections for Gemini 3.1 Flash-Lite, and within range for cost-conscious production workflows.

## 7.2 Test 5: Throughput and Cost

Test 5 was designed to detect emails signaling the sender's intent to keep a topic out of the written record (a theme that exercises implicit intent and context rather than keyword patterns). The test ran on 1,083 emails from the Podesta corpus for March 2016, with average message size around 81 KB.

**Important caveat on accuracy.** We were unable to develop a prompt that worked reliably across all tested models for this task, and classification results are therefore not published for Test 5. Two factors contributed. The Test 5 theme is an absence-detection theme (inferring what an email deliberately is not saying), which is systematically harder for current AI models than presence-detection. Independently, broader testing on the Podesta corpus indicates that subjective themes resist stable inter-model agreement on this corpus regardless of the presence-or-absence framing: a parallel experiment on the presence-detection theme "Damage Control and Crisis Communications" produced Responsive counts varying by a factor of 3.7 across seven models, with 45% of flagged emails representing single-model opinions and only 24 emails (out of a 326-email pooled union) receiving unanimous Responsive verdicts. Prompt-model fit on subjective themes can therefore vary substantially across model families. Detailed findings, including pairwise inter-rater agreement, are documented in the Podesta Corpus Benchmark Methodology Note. Practitioners should treat this as a methodological reminder: some classification tasks require significant prompt-development effort, and prompt validation on representative data is essential before committing to a production run.

With that caveat, Test 5 yielded useful operational data on throughput and cost at production scale. Cost figures below were calculated using Test 1 pricing.

| Model | Duration | Emails / weekend (est.) | Est. weekend cost |
|---|---|---|---|
| Ministral 3 14B | 4m 35s | ~879,000 | $0 (offline) |
| Mistral Small 3.2 24B | 6m 34s | ~614,000 | $0 (offline) |
| Gemini 3.1 Flash-Lite (Agent Platform) | 10m 29s | ~384,000 | ~$165 |
| Gemini 3.1 Flash-Lite | 13m 55s | ~289,000 | ~$123 |
| Gemini 3 Flash (preview) | 14m 47s | ~272,000 | ~$323 |
| OpenAI GPT-5.4 | 18m 21s | ~219,000 | ~$841 |
| Gemma 4 26B | 1h 28m 42s | ~45,500 | $0 (offline) |
| Llama 3.3 70B | 2h 8m 42s | ~31,300 | $0 (offline) |

Test 5 was conducted before Claude 4.7 Opus, Grok 4.2 Non-Reasoning, Qwen 3.6 27B Dense, and Qwen 3.6 35B MoE entered the benchmark. For these four models only Test 1

throughput is currently available; the corresponding Test 1 weekend extrapolations are roughly 300,000 (Grok 4.2 Non-Reasoning), 104,000 (Claude 4.7 Opus), 34,600 (Qwen 3.6 35B MoE), and 18,700 (Qwen 3.6 27B Dense). Production-payload (Test 5) behavior has not been directly measured for these four models.

Per-email token volumes on Test 5 versus Test 1 are model-dependent rather than uniformly higher: approximately +2% to +4% for retained offline models, −5% to −6% for Gemini 3.1 Flash-Lite in both deployments, +16% for OpenAI GPT-5.4, and +50% for Gemini 3 Flash. Payload size affects local-model throughput more than cloud-model throughput, which is bound by API round-trip latency rather than token count.

**Practical takeaways from Test 5:**

- Production cost of classifying a mailbox-scale corpus (hundreds of thousands of emails) with the cheapest cloud model (Gemini 3.1 Flash-Lite, AI Studio deployment) is around ~$42 per 100,000 emails—well into the low hundreds of dollars for a full weekend run.
- OpenAI GPT-5.4 at production scale has a per-email cost roughly 9× that of Gemini 3.1 Flash-Lite (AI Studio), and processes fewer than 80% as many emails per weekend; the combined effect is a weekend cost roughly 7× higher.
- Offline Ministral 3 14B can process nearly 900,000 emails per weekend at zero marginal cost, provided the task is amenable to its capability range.

---

# 8. How AI Classification Compares to Keyword Search and TAR

This section frames the benchmark results against published performance ranges for traditional keyword search and Technology-Assisted Review. The figures for keyword and TAR performance below are drawn from the peer-reviewed literature and standardized evaluation tracks (TREC Legal Track, TREC Total Recall Track), synthesized in the companion document Quantitative Performance Benchmarks for Keyword Search and Technology-Assisted Review in eDiscovery and Digital Forensics (Fookes Software, 2026).

## 8.1 Published Baselines

| Method | Typical Recall | Typical Precision | Typical F1 |
|---|---|---|---|
| Keyword / Boolean search (production) | 20–40% | 10–79% (query-dependent) | ~25–40% |
| TAR 1.0 (SPL / SAL) | 50–75% | 60–80% | ~55–75% |
| TAR 2.0 (CAL) | 75–96% | 80–96% | ~75–96% |
| Human linear review (baseline) | 49–54% | 18–20% | ~27–28% |

Two points from this table are worth underlining:

- **Keyword search misses 60–80% of responsive documents** in typical production use. The Blair & Maron (1985) field study—still the most cited benchmark in the literature—found that experienced legal professionals achieved only 20% recall while

believing they had reached 75% or better. TREC Legal Track reference Boolean queries ranged from under 4% to 24% recall depending on judgment regime.

- **CAL-based TAR 2.0** is the best-evidenced production methodology, reaching 75–96% recall at 80–96% precision in multiple controlled evaluations on large email collections.

## 8.2 Where AI Classification Sits

The models retained in this benchmark all scored above 94% F1 on Test 1, with the top five between 98.8% and 99.6%. Placed against published baselines:

- Every tested model exceeds the entire typical F1 range for keyword search by a wide margin.
- Every tested model meets or exceeds the lower bound of the TAR 2.0 / CAL range, and the top performers match or exceed CAL's upper bound.
- Recall in particular is strong: every model in the set achieved 99%+ recall on Test 1, meaning miss rates below 1% at the operating point tested. This compares favorably against the 4–25% miss rate typical of TAR 2.0, the 25–50% miss rate typical of TAR 1.0, and the 60–80% miss rate typical of keyword search.

One caveat applies: the 6% prevalence in Test 1 is favorable relative to very-low-prevalence investigations, where absolute false-positive counts can still be large despite low false-positive rates.

The corpus design itself reinforces, rather than qualifies, the comparison. The 120 synthetic responsive emails in Test 1 were deliberately written to defeat keyword detection—using paraphrase, indirection, and euphemism in place of the trigger terms a Boolean search query would catch. The intent was to test whether AI models would also be defeated by the vocabulary-mismatch tactics that routinely break keyword search in real investigations. They were not. Every retained model achieved 99%+ recall and 94%+ F1 on this corpus, which is direct evidence of AI's advantage on precisely the language patterns where keyword approaches fail. The test was not handicapping keyword search—it was probing whether AI shares its blind spot. The answer is that it does not.

## 8.3 Practical Implications

The differences between methods translate into concrete operational consequences:

- **Review effort.** TAR 2.0 typically reduces review volume by 50–80% relative to exhaustive manual review or broad keyword culling. AI classification of the kind tested here allows binary triage passes that surface responsive emails for focused review without requiring the per-matter training signal that TAR 2.0 depends on. For matters where TAR 2.0 cannot be easily bootstrapped (small corpora, early-stage investigations, limited reviewer time), AI classification provides a viable alternative with comparable or better accuracy.
- **Risk of missed evidence.** Keyword search leaves 60–80% of responsive emails undisclosed on a typical matter. For investigations where false negatives carry legal or investigative risk (regulatory compliance, FOIA responses, internal misconduct reviews, insider-threat cases), this miss rate is a material liability. AI classification at the performance levels documented here reduces that risk by an order of magnitude.
- **Over-collection.** Keyword search often produces review sets where 70–90% of flagged documents are false positives. The top-performing AI models in this

benchmark produced precision of 90–99%, meaning reviewers spend far less time rejecting irrelevant material.

- **Vocabulary mismatch immunity.** Keyword search fails systematically when authors use informal, coded, or euphemistic language ("the unfortunate situation" for "the accident"). AI models operate on meaning rather than exact strings and are largely immune to this failure mode. This is particularly relevant for insider-threat and corruption investigations, where principals routinely avoid direct language.

### 8.4 Where AI Classification Does Not Exceed Traditional Methods

Three caveats deserve to be stated plainly:

1. **AI classification depends heavily on prompt quality and prompt–model fit.** Test 5 in this benchmark illustrates the point: a task that could not be reliably specified produced wildly divergent results across models on the same corpus, which is why the classification results for that test have been withheld. The difficulty there is intrinsic to the theme—a well-engineered Boolean search query would likely fare no better, and probably worse, on the same task—but the divergence across models underscores that prompt–model fit is a real variable in production deployments. Practitioners deploying AI classification in production should expect to validate prompts on representative data before relying on them for sanction-bearing decisions.
2. **The 65% inter-assessor agreement ceiling applies equally.** Human experts reviewing the same corpus agree on relevance only about 65% of the time (Voorhees, 2000). Apparent errors in AI classification may partly reflect inconsistency in the gold standard itself—a constraint that affects keyword search, TAR, and AI alike.
3. **Absence-detection and highly contextual themes remain hard.** Detecting what an email does not say (Test 5's theme) is systematically more difficult than detecting what it does say. On such themes, even frontier models produce results that vary substantially with prompt formulation.

## 9. Choosing a Model for Your Workflow

The "best" model depends on the specific constraints of the matter at hand. The guidance below maps common Aid4Mail use cases to the most suitable model from the benchmark set.

### 9.1 Large-Volume Binary Classification (Cloud)

*Example workflow: 200,000-email mailbox triaged for insider-threat indicators in a regulatory matter.*

- **Primary choice:** Gemini 3.1 Flash-Lite (F1 96.0%, AY 99.45% for AI Studio / 99.50% for Agent Platform, 1.30 emails/s on Test 5 for AI Studio / 1.72 emails/s for Agent Platform, ~$42 per 100K Test 5 emails for AI Studio / ~$43 for Agent Platform)—the lowest-cost cloud option in the benchmark that still delivers strong accuracy and decision coverage. Use the Agent Platform deployment when the project is eligible for the US or EU multi-region endpoint and speed matters; use AI Studio otherwise.

- **Strong alternative:** Grok 4.2 Non-Reasoning when the workflow can absorb roughly 4× the per-email cost in exchange for the highest cloud F1 and tied-best Automation Yield in the benchmark (99.2% / 99.85%).

## 9.2 Multi-Category Triage or Broad Compliance Monitoring (Cloud)

*Example workflow: initial triage of a newly acquired corpus across five misconduct themes, or ongoing compliance monitoring across issue types.*

- **Primary choice:** Gemini 3.1 Flash-Lite (100% on Test 2, 100% on Test 3, 100% on Test 4 in both AI Studio and Agent Platform deployments) at the lowest cloud price point.
- **Strong alternative:** Gemini 3 Flash when prompt complexity or length justifies the larger model—it also posts a perfect 100% on Tests 2, 3, AND 4, joining Flash-Lite and Qwen 3.6 27B Dense in that group.

## 9.3 High-Stakes Matter with Premium Analysis Needs (Cloud)

*Example workflow: small corpus where every decision is carefully scrutinized, summarization and reasoning are needed alongside classification.*

- **Primary choice:** Claude 4.7 Opus (low effort) for analysis-heavy work that will benefit from its broader capabilities, including a perfect 100% on the Korean multi-category test; combine with a cheaper classifier (Gemini 3.1 Flash-Lite or Grok 4.2 Non-Reasoning) for the bulk classification pass.

## 9.4 Air-Gapped or Data-Residency-Constrained Deployment

*Example workflow: government investigation, classified data, contractual data-residency obligations, or organizational policy prohibiting cloud processing.*

- **Primary choice (multilingual, accuracy-first):** Qwen 3.6 27B Dense—the only offline model in the perfect-on-Tests-2/3/4 group (alongside cloud entries Gemini 3 Flash and both deployments of Gemini 3.1 Flash-Lite), and an F1 of 98.77% on Test 1 with 99.80% AY. Best suited to small or medium corpora where accuracy outweighs throughput.
- **Faster sibling:** Qwen 3.6 35B MoE for higher Test 1 F1 (99.17%) at the same 99.80% AY and roughly 2× the throughput of the Dense variant, with a small accuracy concession on Tests 2/3/4.
- **Balanced multilingual on smaller GPUs:** Gemma 4 26B on a workstation with at least 24 GB of GPU memory. Strong multi-category and multilingual accuracy in a smaller VRAM footprint than the Qwen 3.6 family.
- **Alternative for binary-task workloads:** Mistral Small 3.2 24B on 24–32 GB VRAM—the highest Test 1 F1 in the benchmark on decided emails (99.6%) at strong offline throughput. Best suited to workloads where a small INCONCLUSIVE review queue is acceptable.
- **Alternative for maximum English F1 on appropriate hardware:** Llama 3.3 70B, provided the workstation has at least 80 GB of VRAM to run it efficiently. Llama 3.3 70B also posts the highest Automation Yield of any retained offline model (99.85%), but its Korean handling is weaker than the Qwen 3.6 and Gemma 4 alternatives.

### 9.5 High-Throughput Triage on Modest Hardware

*Example workflow: small forensic practice or individual investigator, limited hardware budget, binary classification dominates the workload.*

- **Primary choice:** Ministral 3 14B on 16 GB VRAM. The fastest model in the benchmark at any price (3.94 emails/s on Test 5), with competitive F1 on binary tasks. Automation Yield of 99.30% on Test 1 confirms the model decides broadly enough to make full use of its throughput advantage.

### 9.6 Decision Matrix

The matrix below summarizes the above guidance by decision dimension.

| Constraint | Primary recommendation |
|---|---|
| Cloud OK, cost-sensitive, binary task | Gemini 3.1 Flash-Lite (Agent Platform if eligible for US/EU regions; AI Studio otherwise) |
| Cloud OK, highest accuracy & coverage | Grok 4.2 Non-Reasoning |
| Cloud OK, multi-category or multilingual | Gemini 3.1 Flash-Lite (perfect on Tests 2/3/4 in both deployments) |
| Cloud OK, premium analysis needs | Claude 4.7 Opus (low effort) |
| Air-gapped, multilingual accuracy-first | Qwen 3.6 27B (Dense) |
| Air-gapped, balanced multilingual | Qwen 3.6 35B (MoE) or Gemma 4 26B |
| Air-gapped, binary-task priority | Mistral Small 3.2 24B |
| Air-gapped, maximum English F1 (80 GB+ VRAM) | Llama 3.3 70B |
| Air-gapped, modest GPU (16 GB VRAM) | Ministral 3 14B |

## 10. Limitations and Caveats

This benchmark is a snapshot of a specific test set against a specific set of models at a specific moment in time. The following limitations apply:

1. **Synthetic responsive content.** The 120 responsive emails in Test 1 and all 200 in Test 2 were synthetic, designed to exercise AI's discrimination advantage over keyword search. Real-world responsive content may exhibit different lexical and structural properties. The Podesta corpus used as the unresponsive background is real and representative, but the responsive signal is not.
2. **6% prevalence in Test 1.** This corresponds to the moderate-prevalence range where TAR 2.0 precision benchmarks are most reliable. In very-low-prevalence investigations (under 1% relevant), observed precision may be lower for all methods, AI included, because of base-rate effects.
3. **Test 1 corpus size-filtered to 1–68 KB.** The 2,000-email Test 1 corpus was filtered to mid-sized emails so that all evaluated models could be benchmarked within a reasonable total runtime. This filter is neutral for accuracy measurement (F1,

precision, recall). Speed, weekend throughput, and cost figures in this report are drawn from Test 5 (1,083 real Podesta emails with no size filter, averaging 81 KB per email) where available, which is likely on the upper end of real-world payload distributions; for models measured in both tests, Test 5 throughput was equal to or higher than Test 1 in seven of eight deployment-rows. The four most recently added models (Claude 4.7 Opus, Grok 4.2 Non-Reasoning, Qwen 3.6 27B Dense, Qwen 3.6 35B MoE) currently have only Test 1 throughput and cost figures, and production-payload behavior has not been directly measured for them.

4. **Single hardware configuration.** Offline-model throughput is hardware-dependent. The RTX 5090 with 32 GB VRAM used here is capable but not at the top of consumer hardware; offline results on smaller GPUs (e.g., 16 GB or 24 GB) may be slower and sometimes less accurate for models that cannot be kept fully GPU-resident.
5. **Prompt-model fit is real.** The same prompt does not perform identically across models. Practitioners should validate their chosen prompt on their chosen model and representative data before deploying to a matter of consequence. Test 5 in particular illustrates how divergent results can be when prompt and model do not mesh.
6. **Cloud-model pricing and capability change.** The cost figures and model-capability assessments reflect pricing and model behavior at the time of testing. Providers adjust pricing, introduce new models, and retire older ones regularly. Notably, both Grok 4.1 Fast and Grok 4.1 Fast+Reasoning (previously included in this benchmark) are scheduled for retirement on May 15, 2026 and have been replaced in the analysis by Grok 4.2 Non-Reasoning. Consult current pricing and model availability before making procurement decisions.
7. **Test 5 classification results not published.** The Test 5 prompt did not perform reliably across all tested models, and classification accuracy results are therefore not reported. Test 5 data is used only for throughput and cost analysis in this document.
8. **Single-matter generalization risk.** Benchmark results should be treated as range indicators, not guarantees. Performance on a specific matter depends on the matter's particulars.

---

# 11. Conclusions

The evidence in this benchmark supports three conclusions that should shape how Aid4Mail users think about AI classification:

1. **AI classification has matured to the point where it reliably outperforms keyword search on realistic forensic and eDiscovery tasks and matches or exceeds the top of the published TAR 2.0 range.** Every retained model in this benchmark achieved F1 above 94% on a 2,000-email corpus with realistic adversarial content, with the top five models clustering between 98.8% and 99.6%.
2. **No single model is best for every task. The right choice depends on the workflow.** A matter in a strict data-residency environment has different requirements than a 500,000-email triage run on a routine regulatory matter. The newest entrants—Qwen 3.6 27B Dense and 35B MoE on the offline side, Grok 4.2 Non-Reasoning and Claude 4.7 Opus on the cloud side—reshape the trade-offs at the top of the table, with Qwen 3.6 27B Dense in particular delivering perfect multilingual and multi-category accuracy at offline throughput. This report's decision matrix summarizes the principal trade-offs; practitioners should match them to their own constraints.

3. **Prompt quality and task nature still matter enormously.** AI classification is not a substitute for investigative judgment; it is a force multiplier for it. A well-specified prompt on a tractable task will produce reliable, defensible results across most models in this benchmark. An ill-specified prompt on a difficult task will produce unreliable results regardless of model choice. Practitioners should treat prompt development as a distinct methodological step with its own validation requirements.

Aid4Mail provides access to all of the models in this benchmark, and the configuration system lets users switch between them without changing anything else in their workflow. The practical implication is that practitioners can select the model best suited to the matter at hand, run a small validation pass on representative data, and proceed with confidence if the results match expectations.

---

*Date of publication: May 15, 2026.*