Aid4Mail™ 6

Aid4Mail AI Email Review Workflow Guide

User Guide

Fookes Software Ltd
Charmey, Switzerland
www.aid4mail.com

Table of Contents

Aid4Mail AI Email Review Workflow Guide

Using AI filtering, classification, and analysis in digital forensics, eDiscovery, FOIA/public-records, litigation support, and internal investigations. **Applies to:** Aid4Mail Investigator and Aid4Mail Enterprise.

Audience: Digital forensics, DFIR, eDiscovery, FOIA/public-records, litigation-support, and internal-investigation professionals who are technically capable but do not need to become AI specialists.

Scope: This guide explains how to use Aid4Mail's AI features in real matters: when to use them, how to choose a deployment path, how to run a first job, how to validate results, and how to preserve defensibility. Detailed provider setup, model-profile tables, benchmark methodology, and advanced local-model tuning belong in companion documents.

1. What Aid4Mail AI Does

Aid4Mail integrates AI into the email processing workflow so that reviewers and investigators can evaluate emails by meaning, not just by exact words or metadata.

The practical goal is simple: use AI to reduce missed evidence, reduce review volume, classify emails more consistently, and support faster triage without requiring a separate predictive-coding project.

Aid4Mail AI can perform three main tasks.

Task	What it does	Typical output
Filter	Decides whether an email meets a defined condition.	Include/exclude from downstream processing.
Classify	Assigns each email to one category or issue label.	Folder names or classification fields in exported output.
Analyze	Produces summaries, translations, extracted fields, or short analytical outputs.	Added fields in PDF, HTML, CSV, TSV, XML, or JSON output.

1.1 Practical examples

Filtering example

> *Flag emails where a participant discusses moving company files to a personal account, private storage service, external device, or unauthorized third party.*

Classification example

> *Classify each email as one of:* `Financial`, `Corruption`, `Discrimination`, `Compliance`, `Exfiltration`, `Clean`, *or* `INCONCLUSIVE`.

Analysis example

> *Summarize the email in two sentences and extract any dates, people, organizations, and locations mentioned by the sender or recipients.*

1.2 What makes this different from generic AI tools

Aid4Mail AI is not a chat interface pasted onto email data. It applies AI inside the email processing pipeline, where each AI output remains tied to a specific source email, source folder, headers, body text, and optional attachment text or metadata.

This matters because forensic and eDiscovery work is not just about producing plausible answers. It is about traceable decisions. Each AI result should be reviewable, exportable, and preservable as part of the matter record.

1.3 What Aid4Mail AI is not

Aid4Mail AI does not replace legal judgment, examiner judgment, or quality control. It should be used as an automated classification and triage layer that supports review, not as an unreviewed final legal conclusion.

AI models can still make classification errors, especially where email content is ambiguous, prompts are vague, categories overlap, or the corpus differs from the data used during testing. For that reason, every production workflow should include sample testing, prompt validation, and preservation of the final output.

2. When to Use AI Instead of Keywords or TAR

AI is most useful when the decision depends on meaning, context, intent, or language rather than exact words.

Keyword search remains valuable. TAR remains valuable in some mature eDiscovery workflows. The question is which tool fits the matter.

2.1 Quick comparison

Need	Keyword search	TAR / predictive coding	Aid4Mail AI
Find a known person, date range, domain, or exact term	Strong	Usually unnecessary	Usually unnecessary
Find concepts expressed with synonyms or indirect language	Weak to moderate	Strong after training	Strong immediately with a good prompt
Start quickly on a small or urgent matter	Strong for simple searches	Often too much setup	Strong fit
Requires seed-set training	No	Yes	No
Change criteria mid-project	Easy, but lexical	Requires retraining or protocol changes	Edit the prompt and retest
Classify into multiple issue categories	Limited	Possible, but usually specialized	Strong fit
Translate, summarize, or extract fields	No	No	Strong fit
Process multilingual email	Requires multilingual query design	May require language-specific workflows	Strong fit with model validation
Keep all AI processing on premises	Yes	Depends on platform	Yes, with offline models

2.2 Use standard Aid4Mail filters first when the rule is simple

Do not use AI to do work that deterministic filters already do faster and at no AI cost.

Use standard Aid4Mail filtering for criteria such as:

- Date ranges.
- Known senders or recipients.
- Domains.
- Folders.
- File types.
- Exact keywords or simple Boolean logic.

Inefficient AI prompt

> *Return `True` if the email was sent between April and June 2024 and exchanged between jane.doe@example.com and joe.doe@example.com. Return `False` otherwise.*

That is better handled as a normal search/filter query.

2.3 Use AI when the issue is semantic

AI is a better fit for questions such as:

- Does this email suggest unauthorized data movement?
- Does the sender appear to be avoiding normal reporting channels?
- Is the discussion about a potential compliance breach or just routine operations?
- Is a forwarded article being merely shared, or did a participant add substantive commentary?
- Does the email contain potential harassment, retaliation, bribery, collusion, or misuse of confidential information?
- Does the email require translation, summarization, or extraction of people/dates/locations?

2.4 When TAR may still be appropriate

TAR may still be preferred when an organization already has mature TAR protocols, review staff, statistical validation practices, and case law experience around a specific production workflow. It may also remain useful for very large productions where the review process is already built around continuous active learning.

Aid4Mail AI is strongest where teams need immediate semantic classification without seed-set training, particularly in forensic triage, internal investigations, smaller or urgent matters, multilingual collections, issue coding, FOIA/public-records review, and matters where offline processing is required.

---

3. Choose Your Deployment Path

Aid4Mail supports three broad AI deployment paths:

1. **Direct cloud API**
2. **Enterprise cloud platform**
3. **Offline local AI model**

Do not treat these as a maturity ladder where cloud is the default and offline is the fallback. For many professional users, a workstation with a 24 GB or 32 GB GPU is within normal equipment budget, and offline AI may be the preferred deployment even when cloud processing is technically allowed.

The right choice depends on data sensitivity, hardware availability, context-window needs, language coverage, throughput, provider policy, and the expected life of the workflow.

3.1 Deployment decision table

Requirement or preference	Strong starting point	Why
Data must not leave the organization	Offline local model	Processing remains on local hardware.
Air-gapped or highly sensitive matter	Offline local model	No cloud API key or internet connection is required once configured.
You already have, or can justify, 24–32 GB VRAM hardware	Offline retained model	Offline AI is a practical first option, not a last resort, when suitable hardware is available.
Zero marginal provider cost after hardware	Offline local model	No per-token cloud charge, though hardware, electricity, and IT support still matter.
Long-term control over a specific model version	Offline local model	A retained local model can continue to be used after a cloud provider retires or replaces an API model.
Fastest setup with no hardware purchase	Direct cloud API	Usually requires only an API key and provider account.
Attachment-heavy analysis requiring more context than local hardware can run efficiently	Large-context cloud model or high-VRAM offline setup	Use cloud when the needed context exceeds practical offline VRAM limits. See Context window limits for offline defaults and tuning.
Corpus primary language or language mix is not supported by available offline models	Language-validated cloud model	Choose a cloud model when the relevant language cannot be handled reliably by the offline models you can run.
Highest precision under a short deadline	Compare cloud and offline	Cloud can be much faster at the highest precision tier; offline may still win if a fast retained model fits the task.
High-volume cloud processing with managed quotas	Enterprise cloud platform	Better suited to sustained batch processing than consumer-style API tiers.
Regional deployment / compliance controls	Enterprise cloud platform	Supports region selection and enterprise controls.
Very high local throughput	Fast offline model	The fastest tested offline model exceeded cloud throughput on the reference workstation.

3.2 Cloud and offline trade-offs

Cloud models have several practical advantages:

- **No high-end local hardware is required.** This matters for occasional users, distributed teams, and organizations that cannot procure or support GPU workstations.
- **Large useful context windows are easier to obtain in the cloud when needed.** Retained cloud models generally provide 200K-class or larger context windows, and several provide 1M or more depending on provider tier and deployment path. These

larger windows are useful when extracted attachment text genuinely requires them, but they rarely improve classification accuracy by themselves. Offline context length is constrained mainly by VRAM; use the operational defaults in Context window limits.

- **Language coverage is a selection constraint, not a cloud default.** Choose a model that has been validated for the corpus's primary language or languages. If an available offline model handles the primary language reliably and the data should remain local, offline remains a sound starting point. Choose a cloud model when the relevant language or language mix is not supported, not validated, or not practical with the offline models and hardware available.
- **They can be much faster at the highest precision tier.** In Test 1, two models tied for the highest Automation Yield: Llama 3.3 70B offline and Grok 4.2 Non-Reasoning in the cloud. Llama took nearly four hours to classify 2,000 emails on the reference workstation, while Grok completed the same task in under 25 minutes—about 9.6 times faster. Llama 3.3 70B requires 80+ GB VRAM for full-speed GPU residency; the comparison would narrow on appropriate high-end hardware.

Cloud models also have five practical disadvantages compared with offline models:

- **Matter data is sent to an external provider or enterprise cloud platform for analysis.** This may be unacceptable for privileged, regulated, confidential, or highly sensitive collections.
- **Enterprise setup can be complex and time-consuming.** Direct API providers are typically simpler, but enterprise platforms require region selection, service accounts, IAM roles, quota configuration, and platform-specific error handling.
- **Rate limits and connection interruptions can affect long jobs.** Enterprise platforms make quotas more manageable, but they do not eliminate operational limits.
- **Cloud use incurs token charges.** For many matters this cost is insignificant compared with manual review, but it is still a usage-based cost that must be approved and monitored.
- **Cloud models have limited service lives.** When a provider retires or changes a model, the exact prior model may no longer be available. Offline models provide more control over long-term repeat use.

3.3 Direct cloud APIs

Direct cloud APIs are the simplest cloud path. You create an account with the AI provider, generate an API key, configure that key in Aid4Mail, and pay the provider directly for usage.

Typical direct API providers include Anthropic, Google, Mistral AI, OpenAI, xAI, and other supported providers.

Direct APIs are practical for evaluation, small to medium jobs, teams that do not need formal enterprise controls, and users who need a context window or language capability that available local hardware and offline models cannot provide. They can be less suitable for sustained high-volume processing if rate limits, quotas, data-residency requirements, or provider terms become operational constraints.

3.4 Enterprise cloud platforms

Enterprise cloud platforms provide managed AI access with stronger operational controls. Examples include Amazon Bedrock, Google Vertex AI / Gemini Enterprise Agent Platform, and Microsoft Foundry.

For Google cloud deployments, treat Google Vertex AI / Gemini Enterprise Agent Platform as one Google Cloud platform family, not two separate enterprise alternatives. For Gemini 3.1 Flash-Lite, Aid4Mail surfaces the `eu` and `us` deployments; the `eu` multi-region endpoint excludes the UK and Switzerland. Do not rely on Gemini 3.1 Flash-Lite for UK or Swiss data residency. Use an offline model or another approved provider/model where that residency is required.

They are generally preferable when you need:

- Regional deployment.
- Enterprise account controls.
- Higher or more predictable quotas.
- Stronger compliance alignment.
- Centralized billing and governance.
- Access to multiple model families through one cloud platform.

Enterprise platforms do not eliminate rate limits. They make them more explicit and manageable. For large batch jobs, this is usually preferable to consumer-grade API throttling, but setup can be more complex than direct API use.

3.5 Offline local AI

Offline local AI uses a model running on your own hardware through tools such as Ollama or LM Studio.

Use offline AI when:

- Case data cannot leave your environment.
- Data-sovereignty requirements prohibit cloud processing.
- You need air-gapped operation.
- You prefer a fixed hardware investment to recurring token charges.
- You want continued access to the same local model after a cloud model is retired or changed.
- Your organization can support the necessary GPU hardware.

Offline processing is a major differentiator. It allows Aid4Mail to apply modern AI classification while keeping the email corpus, prompts, and model inputs inside the organization's infrastructure.

Offline processing is not inherently a lower-quality option. In the benchmark, offline models posted the highest Test 1 F1 score and several near-frontier results. The main offline constraints are hardware, practical context length, model fit for the corpus's primary language or languages, and throughput for very large models.

As a practical hardware guide, 24 GB and 32 GB GPUs open useful offline-model options for professional work. Start with conservative context settings and increase them only when validation shows material truncation or attachment-text loss; see Context window limits for recommended ceilings by VRAM tier.

3.6 Responsibility for provider access and costs

Aid4Mail does not provide direct access to commercial AI services. For cloud use, you must obtain credentials from your chosen provider, accept that provider's terms, and manage provider billing directly.

Before using any cloud provider on a real matter, confirm:

- The provider's data-use terms.
- Whether your data may be used for training.
- Region and data-residency options.
- Retention and logging behavior.
- Data Processing Agreement availability.
- Internal approval for sending matter data to that provider.

---

4. Quick Start: Run Your First AI Job

This section is intended for a first practical run. It assumes at least one AI provider or local model has already been configured in Aid4Mail.

For detailed administrator setup, use the separate Aid4Mail AI Provider and Model Configuration guide.

4.1 First-run workflow

```
Select provider/model
        ↓
Choose Filter, Classify, or Analyze
        ↓
Load or write prompt
        ↓
Decide whether to include attachments
        ↓
Run a small stratified or deliberately selected validation sample
        ↓
Review the sample output:
   - For Filter: review passed items, rejected items from processing logs, errors,
     and any manual-review queue.
   - For Classify: review Responsive, INCONCLUSIVE, and sampled Unresponsive items.
     Use Classify rather than Filter when a three-label auditable field is
required.
   - For Analyze: review factual accuracy, omissions, extraction completeness,
translation quality,
     unsupported statements, and output-field formatting.
        ↓
Refine prompt and repeat if needed
        ↓
Run the production job
        ↓
Export and preserve results, settings, validation notes, and any reviewer tags or
decision records
```

4.2 Step 1 — Configure or select a provider

Open **App Settings > AI** and confirm that your AI provider is available.

At a high level:

Provider type	What Aid4Mail needs
Direct API provider	API key.
Google Vertex AI / Gemini Enterprise Agent Platform	Region, Google Cloud Project ID, Service Account JSON key file.
Amazon Bedrock	Region and AWS IAM credentials.
Microsoft Foundry	API key and resource name.
Ollama / LM Studio	Local endpoint URL and matching context length.

Treat Google Cloud service-account JSON files as sensitive credentials. Store them securely, restrict file access, and do not commit or copy them into shared matter folders unless access is controlled.

For local providers, confirm that the local AI server is running before processing. Default local endpoints are commonly:

Tool	Default endpoint
Ollama	`http://127.0.0.1:11434`
LM Studio	`http://127.0.0.1:1234`

4.3 Step 2 — Choose the AI task

Configuring an AI task in Aid4Mail involves two locations:

- **Project Settings > AI** is where the model, prompt, attachment setting, and (for Classify) category list are configured. This is the focus of §4.3 through §4.6 below.
- **The Settings tab on the Sessions screen** is where the task is actually wired into the run. For AI Filter, this means enabling the AI filter checkbox; for AI Classify, this means setting a folder structure template that includes `{Classify}` (or adding an `AI.Classify` / `X-AI-Classify` field to the output configuration); for AI Analyze, this means choosing an output format that supports added fields and adding the `AI.Analyze` or `X-AI-Analyze` field.

Configuring one location without the other is the most common first-run mistake: a saved prompt with no enabling control on **Sessions > Settings** produces no AI output, and an enabled task with no configured prompt produces an error. See §5 for the full per-task procedures.

Open **Project Settings > AI**. Aid4Mail provides separate configuration sections for:

- **Filter**
- **Classify**
- **Analyze**

You only need to configure the task or tasks you plan to use.

4.4 Step 3 — Select a model

Select a model that fits your matter. For a first run, choose a default model from the summary in Model Selection at a Glance, but do not assume the cloud path should be tested first.

For most first tests:

- Use a retained offline model if suitable local hardware is available and you want data to remain on premises, avoid provider token billing, or preserve long-term access to a specific model.
- Use a low-cost cloud model if you lack suitable GPU hardware, need a context window larger than your local hardware can run efficiently, need a primary language or language mix that available offline models cannot handle reliably, or need rapid high-precision throughput.
- Use an enterprise cloud platform when the matter permits cloud processing but requires regional controls, managed quotas, or enterprise contracting.
- Use a model validated for the corpus's primary language or languages.

4.5 Step 4 — Load or write a prompt

Stay in **Project Settings > AI**. In the relevant task section (Filter, Classify, or Analyze), use the prompt field to write your own prompt or select **Open** to access the prompt library. The Aid4Mail prompt field requires a single-line prompt; convert the final prompt to a single continuous paragraph before saving it. For defensibility and later editing, preserve both versions:

- A readable source version with line breaks, sectioning, and version notes.
- The normalized single-line Aid4Mail version actually used for the run.

Aid4Mail includes a prompt library organized by task and theme, including Digital Forensics, eDiscovery, and FOIA/Public Records. These prompts are starting points. Always review and adapt the prompt to the specific matter.

If the prompt library does not appear, check whether Windows Controlled Folder Access blocked installation of the prompt files, and inspect the **AI Prompts** subfolder of the Aid4Mail program folder.

After editing a prompt, use **Verify** to check it. Verification uses the selected AI model and may incur a small provider cost.

4.6 Step 5 — Decide whether to include attachment data

Attachments are excluded by default. For a first test, leave full attachment text disabled unless attachment content is likely to affect the classification.

In **Project Settings > AI**, enable **Include attachment data** for the relevant task when attachment content must be sent to the model. Under **Options**, set **Attachment text size limit** when processing document-heavy mailboxes or models with limited context windows.

Leaving the limit blank permits unlimited extracted text per attachment, but the combined payload can still exceed the model context window, causing truncation or attachment-text omission.

Aid4Mail always includes attachment names in the email metadata sent to the model, which can be useful even when full attachment text is not included.

Use the attachment guidance in Attachment Strategy before enabling full attachment extraction for a production run.

4.7 Step 6 — Run a small representative sample

Start with a sample large enough to reveal obvious problems but small enough to review manually. A practical first validation sample is often 100–500 emails, but it should not be merely random unless the matter is high-prevalence and low-risk.

For low-prevalence matters, a random sample may contain few or no responsive items. Supplement random sampling with targeted known-positive or likely-positive examples, likely negatives, and borderline items.

The validation sample should be stratified or deliberately selected to include:

- Known or likely responsive items.
- Likely unresponsive items.
- Borderline or ambiguous items.
- Major custodians, time periods, and folders.
- Representative languages.
- Representative attachment types.
- Forwarded or quoted-thread examples.
- Items likely to trigger privilege, redaction, exemption, or sensitivity handling, if those issues are in scope.

For prompt testing and early validation, consider using **HTML** as the target format with **Include portable email viewer** checked. This lets you open the completed sample immediately in a browser, inspect the AI classification results, and tag emails for follow-up without waiting for a separate review-platform import.

If viewer tags are used for prompt validation, QA, escalation, or review decisions, export and archive the tag file. Browser-local tags alone should not be treated as preserved matter records because they are tied to the browser, machine, and email collection.

4.8 Step 7 — Review and refine

Review the output before running the full job.

If you exported the sample to HTML with the portable viewer included, open the viewer from the completed target folder. Use the Classification column, search, folder filtering, and tagging to inspect Responsive, Unresponsive, and INCONCLUSIVE results quickly.

Look for:

- Missed responsive examples.
- Overbroad false positives.
- Too many INCONCLUSIVE results.
- Category confusion.
- Problems with forwarded or quoted material.
- Problems with attachments or multilingual content.

Revise the prompt if needed, then rerun the sample.

4.9 Step 8 — Run the production job

Once the sample results are acceptable, run the production job.

For long jobs:

- Use deterministic scoping filters such as custodian, date range, folder, file type, known nonresponsive sources, and known exclusion criteria to reduce the corpus first.
- Do not use issue-keyword filters as the sole gate into AI review unless the recall impact has been validated and documented.
- Enable incremental processing where appropriate. Incremental processing is not available when pre-acquisition filtering is enabled; for resumable long runs, prefer post-acquisition filtering when that fits the matter.
- Monitor provider quotas and rate limits.
- Preserve the prompt, model, provider, settings, logs, and output.

4.10 Step 9 — Export and preserve the results

Export the AI classification or analysis results in the target output format. For review-heavy validation runs, HTML with the portable viewer is often the fastest way to inspect and tag results in a browser. For production preservation, choose the format required by the matter and archive the AI output generated at the time of the run.

For AI Filter jobs, preserve Aid4Mail's processing logs and filter decision records, which log filter settings and record which emails passed the filter or were rejected. Archive the prompt, model, provider, settings, run date/time, and any errors or throttling events. If the exported output must carry an auditable label for every email, use Classify or Analyze to export that field before culling.

Do not rely on rerunning a model later to reproduce every result exactly.

5. The Three AI Tasks: Filter, Classify, Analyze

Aid4Mail's AI features are configured by task. The same model may be used for more than one task, but each task should have its own prompt and output design.

5.1 AI Filter

AI filtering determines whether an email should pass through the workflow based on a semantic condition.

Use AI filtering when the criterion requires interpretation.

AI Filter is a pass/fail workflow control. Use it carefully when the result may exclude emails from downstream output. For privilege, exemption, sensitive-data, trade-secret, redaction, or legal-review workflows, filtering is appropriate only if both passed and rejected decisions are preserved and reviewable.

If the workflow needs to route, segregate, preserve, or audit both included and excluded items, prefer AI Classify or an exported AI field instead of using AI Filter alone.

Semantic routing use cases—prefer Classify or exported AI fields for high-stakes workflows

High-stakes routing is usually better handled with AI Classify or exported AI fields when the workflow must route, preserve, and audit both sides of the decision. Semantic candidates include:

- Attorney-client privilege, legal-professional privilege, or work-product emails requiring special handling.
- Emails that may be exempt from disclosure under applicable FOIA, public-records, regulatory, or discovery rules.
- Personal emails that should be excluded or segregated for privacy reasons.
- Emails containing protected personal, medical, financial, child-related, or other sensitive information requiring restricted review.
- Confidential business information, trade secrets, or security-sensitive material requiring escalation before release.
- Messages that need redaction, privilege review, or legal-review routing based on meaning rather than keywords alone.

For high-stakes exclusion decisions, do not rely on AI Filter as unreviewed culling. Aid4Mail logs filter settings and records passed and rejected emails; archive those logs with the matter record. If the matter requires an auditable category or basis visible in exported output, use AI Classify or an exported AI field before culling.

Avoid AI filtering for simple deterministic criteria

- Date range.
- Known sender or recipient.
- Known domain.
- Known folder.
- Exact keyword.
- Attachment extension.

Creating an AI Filter task

1. Go to the **Settings** tab on the **Sessions** screen.
2. Under **Filter**, select **Enable AI filtering**.
3. Configure the AI Filter prompt in **Project Settings > AI**.
4. Test the prompt on a small sample before running the full session.
5. Archive Aid4Mail processing logs and AI Filter decision records for both passed and rejected emails, along with the prompt, model, provider, settings, and run date/time.

5.2 AI Classify

AI classification assigns a label to each email.

Use classification when you want to group emails into review folders, issue categories, or structured export fields.

Open-ended classification

In open-ended classification, the prompt asks the model to determine the category.

Example:

```
Identify the primary language used in this email. Reply only with the agreed
language label, such as English, French, German, Spanish, Italian, or Korean. If
the language cannot be determined reliably, reply INCONCLUSIVE.
```

Restricted classification

In restricted classification, you provide a fixed category list and require the model to choose from it.

Example:

```
Classify this email into exactly one of the following categories:
Financial, Corruption, Discrimination, Compliance, Exfiltration, Clean,
INCONCLUSIVE.
Use INCONCLUSIVE only when the email contains relevant signals but does not provide
enough information to choose a category reliably.
```

Restricted classification is usually easier to audit because every output must fit a known label set.

Creating an AI Classify task

1. Go to the **Settings** tab on the **Sessions** screen.
2. Under **Target**, select **Use a template** from the **Folder structure** list.
3. In **Folder structure template**, insert `{Classify}`.
4. Configure the AI Classify prompt in **Project Settings > AI**.
5. Optionally enter a comma-separated list of allowed categories for restricted classification.

Aid4Mail will create folders named after classification results when the folder structure uses the classification value.

Exporting AI Classify results as a field

Use this path when the classification value must be preserved in PDF, HTML, CSV, TSV, XML, or JSON output rather than only reflected in the target folder structure.

1. Go to the **Settings** tab on the **Sessions** screen.
2. Choose a target format that supports added fields: **PDF**, **HTML**, **CSV**, **TSV**, **XML**, or **JSON**.
3. Open the relevant output configuration editor: **Email header configuration** for HTML/PDF, **Column configuration** for CSV/TSV, or **Content configuration** for XML/JSON.
4. Add the AI classification field to the selected output items. Depending on the editor, this appears as **AI.Classify** for structured data outputs such as CSV, JSON, or XML, or **X-AI-Classify** for target email/header-style outputs such as HTML or PDF.
5. Include other useful fields such as Subject, From, To, Date, source folder, Message-ID, and attachment names.
6. Select the model and configure the AI Classify prompt in **Project Settings > AI**.
7. If using restricted classification, enter the allowed category list.

8. Click **Save** to store the custom prompt and category list for reuse; for defensibility, also archive the final prompt, category list, and processing log for the matter.
9. Run a validation sample and confirm that the exported field contains the expected classification value for each email.

5.3 AI Analyze

AI analysis generates a structured or text-based result from each email.

Use AI Analyze for:

- Summaries.
- Translation.
- Entity extraction.
- Date extraction.
- Issue notes.
- Sentiment/tone indicators.
- Short risk explanations.
- Reviewer-assistance fields.

AI analysis is available for export formats that can include additional fields, such as PDF, HTML, CSV, TSV, XML, and JSON.

Creating an AI Analyze task

1. Go to the **Settings** tab on the **Sessions** screen.
2. Choose a target format that supports added fields: **PDF**, **HTML**, **CSV**, **TSV**, **XML**, or **JSON**.
3. Open the relevant output configuration editor: **Email header configuration** for HTML/PDF, **Column configuration** for CSV/TSV, or **Content configuration** for XML/JSON.
4. Add the AI analysis field to the selected output items. Depending on the editor, this appears as **AI.Analyze** for structured data outputs such as CSV, JSON, or XML, or **X-AI-Analyze** for target email/header-style outputs such as HTML or PDF.
5. Include other useful fields such as Subject, From, To, Date, source folder, Message-ID, attachment names, and classification (**AI.Classify** or **X-AI-Classify**) if AI Classify is also configured.
6. Select the model and configure the AI Analyze prompt in **Project Settings > AI**.
7. Set a maximum output-token limit suitable for the task.
8. Click **Save** to store the custom prompt for reuse; for defensibility, also archive the final prompt and processing log for the matter.
9. Run a validation sample and confirm that the exported field contains the expected classification value for each email.

For analysis tasks, output-token usage can be higher than for filtering and classification. Keep prompts focused.

Validate Analyze outputs differently from Filter or Classify outputs. For summaries, check omissions and unsupported statements. For translation, check key passages in the source language where possible. For extraction tasks, check both false positives and missed entities, dates, people, organizations, and locations. For reviewer-assistance notes or risk

explanations, confirm that the output is supported by the source email and does not introduce facts not present in the email.

---

6. Prompting for Practical, Defensible Results

The prompt is the instruction that tells the model what to decide. In Aid4Mail AI workflows, prompt quality directly affects accuracy, consistency, cost, and defensibility.

A good prompt should be clear enough that a reviewer can understand the rule, a model can apply it consistently, and a later reviewer can audit what was intended.

6.1 Use the three-label pattern for review workflows

For most investigation and review workflows, use three labels:

- **Responsive** — the email clearly meets the defined criteria.
- **Unresponsive** — the email does not meet the criteria.
- **INCONCLUSIVE** — the email contains relevant signals but is too ambiguous for a reliable decision.

This is safer than forcing every email into Responsive or Unresponsive. INCONCLUSIVE becomes the human-review queue.

6.2 Basic prompt structure

A strong classification prompt usually includes:

1. **Role** — who the model should act as.
2. **Task** — what decision must be made.
3. **Responsive criteria** — what must be present.
4. **Unresponsive criteria** — what does not count.
5. **INCONCLUSIVE threshold** — when to defer.
6. **Source limitation** — whether to consider forwarded/quoted content.
7. **Output format** — the exact label or structure to return.

The prompt examples below use line breaks to make the instructions easier to read in this guide. When entering prompts into Aid4Mail, enter them as a single continuous prompt without line breaks. Preserve the same meaning with punctuation, semicolons, and explicit label names rather than relying on visual formatting.

6.3 Example: focused binary prompt

```
You are a digital forensics investigator. Classify the email into exactly one
category.

Reply Responsive if the email shows data exfiltration or insider-threat activity,
such as unauthorized transfer of files or datasets to external parties, use of
personal storage or messaging to bypass monitoring, sharing proprietary information
with competitors or foreign contacts, credential sharing or unauthorized access,
```

```
systematic data collection before departure, threats to delete or corrupt data, or
attempts to circumvent security controls.

Reply INCONCLUSIVE if the content suggests one of these behaviors but is ambiguous
or lacks enough context for a reliable decision.

Otherwise, reply Unresponsive.
```

6.4 Example: prompt with forwarded-content rule

Forwarded content often creates false positives. A participant may forward an article or thread without endorsing or discussing it. Make the rule explicit.

```
When evaluating forwarded or quoted content, base the classification only on
comments written by the actual email participants. Classify the email as
Unresponsive when the relevant language appears only in a forwarded article, quoted
thread, or third-party message and the actual participants add no substantive
commentary of their own.
```

6.5 Example: multi-category prompt

```
You are a digital forensics investigator. Classify the email into exactly one
category:

Financial – illicit financial activity, cryptocurrency fraud, suspicious
transactions, or money laundering.
Corruption – bribery, kickbacks, quid pro quo arrangements, gifts for influence, or
improper attempts to influence officials or regulators.
Discrimination – evidence of employment discrimination, harassment, hostile work
environment, or retaliation based on a protected characteristic.
Compliance – regulatory breaches or internal policy violations that do not clearly
involve financial crime or corrupt influence.
Exfiltration – unauthorized transfer, leakage, or suspicious movement of documents,
files, or datasets.
Clean – none of the above categories apply.
INCONCLUSIVE – relevant signals exist but the content is too ambiguous to classify
reliably.

Reply with only one category name.
```

6.6 Keep categories mutually exclusive

Overlapping categories cause inconsistent outputs. If categories overlap in the real matter, define precedence.

Example:

```
If an email fits both Exfiltration and Compliance, choose Exfiltration.
If an email fits both Financial and Corruption, choose Corruption only when
improper influence or bribery is explicit; otherwise choose Financial.
```

6.7 Define what is not responsive

Many poor prompts overdefine Responsive and underdefine Unresponsive. That causes over-collection.

For example, in a FOIA or public-interest workflow, specify that routine press monitoring, scheduling, campaign operations, newsletters, automated alerts, and forwarded articles without participant commentary are Unresponsive unless the participant-authored text independently meets the Responsive criteria.

6.8 Use INCONCLUSIVE deliberately

INCONCLUSIVE should not be a dumping ground for uncertainty caused by vague prompts. It should mean:

- The email contains a relevant signal.
- The signal is not clear enough for a reliable committed decision.
- A human reviewer or second-pass model should review it.

If too many items are INCONCLUSIVE, the prompt may be too broad, too strict, or internally inconsistent.

If too few items are INCONCLUSIVE and false positives are high, the prompt may be forcing decisions too aggressively.

6.9 Test before production

Before processing a large corpus, define acceptance criteria for production readiness. Record the maximum tolerated missed responsive items in the validation sample, acceptable false-positive burden, acceptable INCONCLUSIVE queue size, second-review requirements, escalation criteria, and named approver.

Then:

1. Build a stratified or deliberately selected validation sample.
2. Include known or likely responsive examples, likely unresponsive examples, borderline items, relevant languages, major custodians, important time periods, forwarded/quoted threads, and representative attachment types.
3. For low-prevalence matters, supplement random sampling with targeted known-positive or likely-positive examples.
4. Run the prompt on the validation sample.
5. Review all Responsive results.
6. Review all INCONCLUSIVE results.
7. Spot-check Unresponsive results, with extra sampling for high-risk custodians, uncommon languages, large attachments, and issue types where missed evidence would be costly.
8. Record the sampling rationale, sample size, model, prompt version, attachment setting, and observed false positives/false negatives.
9. Adjust the prompt and repeat until the result meets the documented acceptance criteria.

For prompt-validation runs, HTML output with **Include portable email viewer** enabled is often the most efficient target. It allows immediate browser-based review of the classification column and lets reviewers tag examples for prompt refinement or escalation.

6.10 Use the prompt library as a starting point

Aid4Mail includes a library of 200+ pre-written prompts across 72 themes, including Digital Forensics, eDiscovery, and FOIA/Public Records themes. If the library is missing, see Prompt library missing.

Use these prompts to avoid starting from a blank page, but do not use them blindly. A prompt that is defensible in one matter may be overbroad or underinclusive in another.

6.11 Document prompt changes

For production runs, preserve:

- Prompt text.
- Prompt version or change history.
- Model and provider.
- Attachment setting.
- Category list.
- Date and time of run.
- Sample-validation results.
- Final exported AI output.

7. Model Selection at a Glance

Model choice affects accuracy, speed, cost, privacy, data residency, and reproducibility. The best model is not always the largest or most expensive one.

This section gives practical recommendations. For complete methodology and per-model analysis, see the separate benchmark report.

7.1 Evidence at a glance

In the May 2026 Aid4Mail AI classification benchmark:

- Eleven model families were retained for headline analysis from a larger benchmark evaluation pool; Gemini 3.1 Flash-Lite is reported as two deployment rows where speed, cost, throughput, or regional behavior matters.
- The retained set included both cloud and offline models.
- Every retained model achieved more than 94% F1 on the primary 2,000-email responsiveness test.
- The top five F1 scores ranged from 98.8% to 99.6%.
- Every retained model achieved at least 99% recall on the primary test.
- Offline models were not second-class: the top F1 result came from an offline model, and several offline models performed at or near frontier cloud-model accuracy.
- The fastest tested offline model processed an estimated 879,000 emails over a 62-hour weekend window on the Test 5 production-payload basis.
- The fastest measured cloud deployment processed about 384,000 emails over a comparable weekend window on the Test 5 production-payload basis.

- The lowest-cost measured cloud deployment was estimated at roughly $42 per 100,000 emails at Test 5 production-payload sizes; the Agent Platform deployment was roughly $43.

Benchmark results do not guarantee performance in a specific matter. Corpus characteristics, prompt quality, prevalence, language mix, attachments, provider limits, and model choice all matter.

7.2 Recommended defaults by priority

Priority	Practical recommendation	Notes
On-premises processing with professional workstation hardware	**Mistral Small 3.2 24B** or another retained offline model	Good first test when 24–32 GB VRAM hardware is available and data control matters.
Strictly air-gapped or no external data transfer	**Ollama or LM Studio with retained offline model**	No cloud provider, no API key, no external processing.
Lowest-cost cloud classification	**Gemini 3.1 Flash-Lite**	Strong multilingual performance and low cost. Use Agent Platform when the matter is eligible for the US or EU multi-region endpoint and speed matters.
Higher cloud precision / decision coverage	**Grok 4.2 Non-Reasoning**	Highest Test 1 F1 among retained cloud models and tied for top Automation Yield; higher cost than Gemini 3.1 Flash-Lite.
Attachment-heavy analysis where extracted attachment text may exceed available offline context	**Cloud model with 200K-class or larger context, or high-VRAM offline setup if local processing is required**	Use cloud when the required context exceeds what your offline hardware can keep GPU-resident. See Context window limits before increasing local context length.
Primary language or language mix not supported by available offline models	**Language-validated cloud model**	Prefer cloud only when the matter's language requirements cannot be met reliably by the offline models and hardware available.
Premium analysis and complex language-heavy review	**Claude 4.7 Opus (low effort)**	Best reserved for high-value analysis, translation, or reasoning-heavy tasks because of cost.
Balanced offline production	**Mistral Small 3.2 24B**	Highest Test 1 F1 on decided emails; strong speed; accepts a small INCONCLUSIVE queue.
Fastest offline triage	**Ministral 3 14B**	Very fast and modest hardware requirements; weaker for broad multi-category classification.
Highest offline multilingual consistency	**Qwen 3.6 27B Dense**	Perfect Tests 2, 3, and 4; the only offline model in that perfect-score group. Qwen 3.6 35B MoE is a close throughput-optimized alternative.

High English-dominant offline decision coverage with high-end hardware	**Llama 3.3 70B**	Strong English classification and near-complete decision coverage; requires approximately 80 GB VRAM minimum for full GPU residency at 64K context and should be validated separately for other primary languages.
Long-term stability of a specific model version	**Offline local model**	Cloud providers can retire or alter models; local models can be preserved under your control.

7.3 Main model trade-offs

Model	Deployment	Best fit	Main caveat
Gemini 3.1 Flash-Lite	Cloud	Low-cost, high-volume, multilingual cloud workflows; Agent Platform when speed or enterprise authentication matters	More false positives than some higher-cost models; Agent Platform `eu` deployment excludes UK and Swiss residency.
Grok 4.2 Non-Reasoning	Cloud	Highest-F1 retained cloud model with top Automation Yield	More expensive than Gemini 3.1 Flash-Lite; Test 5 production-payload throughput was not measured.
GPT-5.4	Cloud	Organizations standardized on OpenAI; analysis use cases	Cost higher than some similarly accurate alternatives.
Claude 4.7 Opus low effort	Cloud / enterprise cloud	Complex analysis, translation, reasoning-heavy tasks	Premium cost; not ideal for routine large-volume classification.
Mistral Small 3.2 24B	Offline	Balanced on-premises classification	Has the highest INCONCLUSIVE count in the retained set (26 on Test 1), which lowers its Automation Yield despite the best decided-emails F1.
Ministral 3 14B	Offline	High-throughput binary triage on modest GPU hardware	Weaker broad multi-category discrimination.
Qwen 3.6 27B Dense	Offline	Accuracy-critical multilingual small or medium corpora	Slowest retained model on the 32 GB reference workstation (~0.08 emails/s, ~6h 38m for 2,000 emails). Not production-viable for weekend-scale batches on 32 GB VRAM; faster hardware required for high-volume use.
Qwen 3.6 35B MoE	Offline	Multilingual accuracy with better throughput than Qwen Dense	Requires more VRAM than 24B-class models.
Gemma 4 26B Think	Offline	Strong Korean/multilingual offline work on 24 GB-class hardware	Slower than Mistral Small and Ministral.

Llama 3.3 70B	Offline	High-accuracy English-dominant work in air-gapped environments	Requires approximately 80 GB VRAM for full-speed operation at 64K context; validate separately for other primary languages.

7.4 Do not choose by accuracy alone

Accuracy is critical, but it is not the only operational metric.

For real matters, also consider:

- Does the model over-collect?
- Does it miss relevant emails?
- Does it produce too many INCONCLUSIVE items?
- Can it handle the languages in the corpus?
- Can it process the required volume by the deadline?
- Is the provider acceptable under privacy and data-residency requirements?
- Is cost predictable?
- Can the result be preserved and defended?

7.5 Cloud versus offline: practical interpretation

Cloud models are convenient, broadly capable, and often fast. They are especially attractive when you lack suitable GPU hardware, need context sizes that your local hardware cannot keep GPU-resident, need support for a primary language or language mix that available offline models cannot handle reliably, or need high-precision throughput under a short deadline.

Offline models are not a last-resort option. They are often the better professional choice when data control, fixed-cost economics, repeatability of the model environment, or independence from provider model retirement matters. With a 24 GB or 32 GB VRAM workstation, several retained offline models become practical for real matters.

Read the trade-off as follows:

Factor	Cloud advantage	Offline advantage
Hardware	No GPU workstation required.	Uses hardware you control; no provider dependency once configured.
Data handling	External processing under provider/platform terms.	Email payloads, prompts, and model inputs remain on premises.
Context length	Retained cloud models generally provide 200K-class or larger context windows, and several provide 1M or more. These larger windows are useful mainly when extracted attachment text genuinely requires them.	Offline context length is mainly a VRAM and GPU-residency trade-off. Start conservatively and see Context window limits before increasing local context length.
Language fit	Useful when the corpus's primary language or language mix is not supported or validated offline.	Strong for validated languages, especially with Qwen, Gemma, and Mistral-family models; offline remains appropriate when the primary language is supported.

Throughput	Can be much faster at the highest precision tier.	Fast retained offline models can exceed cloud throughput on suitable hardware and tasks.
Cost model	Token billing; often small compared with manual review.	Hardware and electricity costs, but no provider token bill.
Operational risk	Rate limits, API errors, outages, and model retirement.	Local setup, GPU capacity, and context tuning must be managed.

A practical selection pattern is to test a retained offline model first when local processing is permitted and the corpus language is supported. Add a cloud comparison when the matter needs more context than the local GPU can run efficiently, requires a language or language mix not validated offline, or demands faster high-precision completion.

8. Cost, Throughput, and Scaling

AI cost and throughput depend on provider, model, corpus size, payload size, prompt length, attachment inclusion, and rate limits.

For Filter and Classify tasks, output tokens are usually minimal. Input tokens dominate cost because each email's headers, body, and optional attachment text must be sent to the model.

8.1 Planning questions

Before a production run, answer these questions:

1. How many emails will be processed after deterministic filtering?
2. What proportion have large bodies or attachments?
3. Will attachment text be included?
4. Which languages are present?
5. Which model will be used?
6. Is the provider direct API, enterprise cloud, or offline?
7. Are quotas sufficient for the deadline?
8. What output must be preserved?

8.2 Use deterministic reduction before AI

To reduce cost and processing time:

1. Apply deterministic pre-acquisition filters where available, such as date ranges, folders, custodians, domains, file types, known nonresponsive sources, and known exclusion criteria.
2. Export or collect the reduced set.
3. Apply post-acquisition filters locally.
4. Collect or include cloud attachments only after the corpus has been narrowed, when possible.
5. Apply AI only to the emails that remain.

Do not use issue-keyword filters as the sole gate into AI review unless the recall impact has been validated and documented. Those filters can exclude the synonym-heavy or indirect-language emails that semantic AI is intended to find.

In a single Aid4Mail run, post-acquisition filtering, cloud-attachment collection, and AI processing can occur in sequence. If incremental processing is enabled, Aid4Mail can resume without starting over after an interruption. Incremental processing is not available when pre-acquisition filtering is enabled; for resumable long runs, prefer post-acquisition filtering when that fits the matter.

8.3 Weekend-throughput reference points

A useful operational benchmark is a 62-hour unattended run from Friday evening to Monday morning.

Path	Example model	Estimated weekend throughput	Measurement basis	Cost implication
Fastest offline	Ministral 3 14B	~879,000 emails	Test 5 production-payload throughput	No provider cost after hardware.
Balanced offline	Mistral Small 3.2 24B	~614,000 emails	Test 5 production-payload throughput	No provider cost after hardware.
Fastest measured cloud	Gemini 3.1 Flash-Lite (Agent Platform)	~384,000 emails	Test 5 production-payload throughput	Provider token cost applies; roughly $43 per 100,000 emails.
Lowest-cost direct API cloud	Gemini 3.1 Flash-Lite	~289,000 emails	Test 5 production-payload throughput	Lowest measured cloud cost; roughly $42 per 100,000 emails.
High-precision cloud estimate	Grok 4.2 Non-Reasoning	~300,000 emails	Test 1-derived estimate; Test 5 not run	Provider token cost applies; highest-F1 retained cloud model.
Premium cloud	Claude 4.7 Opus low effort	~104,000 emails	Test 1-derived estimate; Test 5 not run	High provider cost; use selectively.

Grok 4.2 Non-Reasoning and Claude 4.7 Opus low effort were not part of the Test 5 production-payload throughput run. Their weekend-throughput figures are derived from Test 1, which used a size-filtered corpus. Treat these as planning estimates, not production-payload measurements. Both Gemini 3.1 Flash-Lite deployments were measured on Test 5; Agent Platform finished the Test 5 corpus about 25% sooner than AI Studio, or about 33% higher emails/s.

Treat these as planning figures, not guarantees. Actual throughput varies by provider load, region, payload size, prompt complexity, hardware, and rate limits.

8.4 Cost reference points

At production-payload sizes in the benchmark, Gemini 3.1 Flash-Lite was estimated at roughly $42 per 100,000 emails via Google AI Studio; the Agent Platform deployment was

roughly $43. More expensive cloud models can cost up to about 30 times that amount, depending on token pricing and output length.

Pricing examples are benchmark reference points, not current provider quotes. Verify provider pricing, output-token pricing, model status, regional availability, and quota terms before approving a production run.

Offline models have no per-email provider charge once hardware is in place, but the workstation, GPU, electricity, maintenance, and IT support should still be considered.

Cloud token costs are often small compared with the cost of manual review, but they remain usage-based costs that should be estimated, approved, and monitored.

8.5 Cost is not the same as cost-effectiveness

The cheapest model is not always the cheapest workflow.

A model with lower provider cost may create more downstream review cost if it over-collects. A model with higher F1 but many INCONCLUSIVE results may require more follow-up review. A premium model may be justified for complex analysis but wasteful for straightforward classification.

Evaluate cost together with:

- Precision.
- Recall.
- Automation Yield.
- INCONCLUSIVE volume.
- Reviewer cost.
- Deadline.
- Data-sensitivity constraints.

9. Attachment Strategy

Attachment inclusion can improve results when the evidence is in the attachment. It can also increase cost, slow processing, and add noise when the email body and attachment filename are already sufficient.

Aid4Mail includes attachment filenames (including extensions) in the metadata sent to the AI model even when full attachment text is not included. This often provides useful signal with little cost.

9.1 What can be included

When attachment inclusion is enabled, Aid4Mail can extract and send:

- Text from word-processor documents.
- Text from spreadsheets.
- Text from presentations.
- Text from PDFs.

- Plain text, CSV, and Markdown content.
- Camera metadata from supported image/photo files, such as raw photos, TIFFs, and JPEGs, where metadata is available.
- Nested files from supported cloud attachments or archive attachments, where available and accessible.

Support depends on file type, encryption/password protection, extraction availability, cloud-attachment accessibility, archive structure, and whether the content is text-extractable. Validate attachment extraction on a representative sample before relying on it for production classification.

Aid4Mail cannot extract the contents of encrypted or password-protected files and does not perform OCR to extract text from image-only content. Payload truncation is reported in a log file that lists the EDRM MIH (Message Identification Hash) signature of the affected email.

Do not judge the AI payload by the attachment's file size on disk. A PDF or word-processor document may be several megabytes because the file includes embedded images, fonts, layout data, style definitions, metadata, compression structures, and other document-format overhead. When Aid4Mail includes attachment content, it does not send the original binary attachment to the AI model. It extracts text, or supported metadata, and adds that extracted content to the normalized AI payload. In many cases, the extracted text is much smaller than the source file. The practical constraint is therefore not the attachment's stored file size, but the amount of extractable text added to the per-email payload, together with the email body, selected headers, prompt, task instructions, category list or output schema, and any other included attachments. Text-heavy documents, spreadsheets, and emails with multiple attachments can still exceed the configured context window, so use the **Attachment text size limit** and validate representative samples before production.

9.2 Attachment decision table

Matter type	Include full attachment text?	Reason
Phishing, malware, spoofing, email attack vectors	Usually no	Body text, URLs, headers, and attachment names often carry the signal.
Insider threat / data exfiltration	Often yes for document-heavy matters	Evidence may be in filenames, spreadsheets, PDFs, or transferred documents.
IP theft / trade secret leakage	Often yes	Attachment contents may define the sensitive material.
FOIA / public records	Test first	Some corpora gain little from attachment text; others depend on it.
Non-consensual imagery or image-based evidence	Include photo metadata	Metadata can be high-signal and low-token.
Broad early-stage triage	Usually sample first	Attachment inclusion can increase cost and noise.

Privilege or legal issue review	Often yes, but controlled	Attachments may contain the privileged or responsive substance.

9.3 Production Pilot lesson

In the large-scale Production Pilot, enabling document attachment extraction increased input-token volume by approximately 32%. The benefit was corpus-specific: one model found 11 additional Responsive items, while another found no additional Responsive items and shifted some items from INCONCLUSIVE to Unresponsive.

The practical conclusion is not “never include attachments.” It is: test attachment inclusion on a representative sample before applying it to an entire corpus.

9.4 Context window limits

Every model has a maximum amount of text it can process in one request. This is called the context window.

Context length is one of the clearest operational differences between cloud and offline processing, but larger is not automatically better. Retained cloud models generally provide 200K-class or larger context windows, and several provide 1M or more depending on provider tier and deployment path. These larger windows are useful when extracted attachment text genuinely requires them, but they rarely improve classification accuracy by themselves.

For offline models, context length is mainly a VRAM-management decision. Use 32K as the safest first-run default on 32 GB VRAM systems. Use 32K–64K as the normal production tuning range, increasing context length only when validation shows material truncation or attachment text loss. Higher context settings reserve more KV-cache memory and may force CPU offload.

If an email plus its extracted attachments exceeds the configured context window, content may be truncated before classification. Payload truncation is reported in a log file that lists the EDRM MIH signature of the affected email. Increase context length only to match the matter’s actual payload size; beyond that point, larger windows mainly reserve more KV-cache memory, reduce throughput, and rarely improve classification accuracy by themselves.

Recommended offline context length by installed VRAM

The table below gives conservative practical ceilings for Q4-class retained local models when the goal is to keep the model fully GPU-resident. Treat these as upper limits, not defaults. For production classification, choose the smallest context length that avoids material truncation on the target corpus.

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:

- The table shows conservative GPU-resident operating recommendations, not absolute architectural limits. Practical context length is still constrained by VRAM, KV-cache memory, quantization, active model size, and throughput requirements.
- The 32 GB reference workstation does not support 256K on any retained 256K-capable model. Reaching 256K requires 40 GB VRAM for the 14B row, 80 GB for the 20–27B row, and 96 GB for the 35B MoE row.
- 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.
- Start with the smallest context length that avoids material truncation. On 32 GB systems, 32K is the safest first-run default; 32K–64K is the usual production tuning range.
- If you experience unexpectedly slow processing, try reducing the context length. This is often the single most effective tuning adjustment.
- A dash 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.
- Hardware fit is not the same as production-viable throughput. A model that loads at the listed context length on a given VRAM tier may still be impractical for weekend-scale batches if its emails-per-second rate is low on that hardware. **Qwen 3.6 27B Dense on the 32 GB reference workstation is a notable case**: it fits at 64K context but ran at roughly 0.08 emails/s in Test 1, completing 2,000 emails in about 6h 38m. That is suitable for small-to-medium corpora and accuracy-critical multilingual work, not for production-scale weekend runs. Faster hardware substantially improves this throughput; see the model-selection guide for current recommendations.

9.5 Suggested starting limits for extracted attachment text

Use attachment text-size limits to control payload size. Set this in **Project Settings > AI** under **Options** as **Attachment text size limit**. The values below are starting per-attachment caps, not guaranteed safe total request sizes.

Model context window	Starting extracted-text cap per attachment
~2M context	Up to 200 KB per attachment
~1M context	Up to 150 KB per attachment

200K–256K context	Up to 75 KB per attachment
128K context	Up to 50 KB per attachment
32K context	Up to 20 KB per attachment

These are starting points. For context sizes not listed, start with the lower neighboring tier and validate against your own corpus. Attachment text limits should be set to preserve the text that matters for the review question, not to fill the model's entire available context window.

The cumulative payload must still fit within the configured context window. That payload includes selected headers, Aid4Mail metadata, the email body, the user prompt, fixed/task instructions, output schema or category list, and extracted text from all included attachments. Multiple attachments can exceed the practical budget even when each individual attachment is below the per-attachment cap.

For attachment-heavy matters, validate both per-attachment limits and per-email total payload behavior on a representative sample.

---

10. Privacy, Security, and Data Residency

AI processing raises data-handling questions that must be addressed before production use.

The most important question is: **where does the email content go?**

10.1 What leaves the environment in cloud workflows

For cloud AI workflows, Aid4Mail prepares a normalized message payload and sends it to the selected AI provider or platform.

Aid4Mail reduces payload size by normalizing the message, including operations such as decoding MIME content, converting HTML bodies to plain text, and retaining selected headers.

Data sent to the model typically includes:

- Selected message headers, such as From, To, Cc, Bcc, Subject, Date, Message-ID, Reply-To, Sender, and related fields.
- Aid4Mail-specific metadata such as source folder, attachment names, Gmail labels, MAPI categories, and status flags.
- The email body converted to plain text.
- Optional extracted attachment text or metadata, if attachment inclusion is enabled.
- The user prompt, including any matter-specific criteria, legal theories, category definitions, examples, or reviewer instructions.
- The category list for restricted classification, if used.
- Aid4Mail's task instructions, validation instructions, output schema, and formatting requirements needed for the selected AI task.
- Runtime settings needed by the provider request, such as model, output-token limit, and temperature handling where supported.

Because prompt text and category labels can reveal sensitive matter strategy, they should be reviewed under the same privacy, privilege, and confidentiality standards as the email content itself.

10.2 What stays on premises in offline workflows

For offline workflows using Ollama or LM Studio, AI processing runs on your organization's hardware. No cloud API key is required for the AI model, and no email payload needs to be sent to an external provider.

This is the preferred path when:

- The matter is highly sensitive.
- Data cannot leave the organization.
- Data sovereignty rules apply.
- The environment is air-gapped.
- Provider terms are unacceptable.

10.3 Enterprise cloud controls

Enterprise cloud platforms may help satisfy operational and compliance requirements by supporting regional deployment, stronger account governance, quota management, and enterprise contracting.

Before using enterprise cloud AI, confirm:

- The selected model is available in the required region.
- Provider terms meet matter requirements.
- A Data Processing Agreement is in place where required.
- Logging and retention settings are understood.
- Internal security has approved the workflow.
- Service-account JSON files and API keys are stored as sensitive credentials, with access restricted to authorized users.
- Any cross-border transfer mechanism is documented.

10.4 Legal and privacy principles

This guide does not provide legal advice. Consult legal counsel and privacy/security teams for matter-specific decisions.

For most matters, document the following:

- Legal basis for processing.
- Data minimization steps.
- Provider and platform used.
- Region used, if cloud-based.
- Data-transfer mechanism, if applicable.
- Prompt and AI settings.
- Attachment setting.
- Retention and deletion plan.
- Chain-of-custody handling.

11. Multilingual Review

AI can reduce the burden of multilingual email review because the model evaluates meaning rather than exact language-specific keywords.

This does not mean every model performs equally well in every language. Model choice still matters, and multilingual corpora should be tested before full-scale processing.

11.1 Practical guidance

Use multilingual AI workflows when:

- The corpus contains multiple languages.
- Translating the entire corpus before triage is impractical.
- Keyword searches would require many language-specific synonym sets.
- The review question is semantic rather than lexical.

11.2 Benchmark-supported language observations

The benchmark directly tested English, French, Spanish, and Korean content. All retained models supported multilingual classification at some level, and the best-performing models handled Korean binary and multi-category tests at very high accuracy.

German and Italian support for the retained offline set is based on model language-support expectations and should be validated on the target corpus before production use. Offline may also be appropriate for other languages when a suitable model is available and the validation sample performs well. Choose a cloud model when the corpus's primary language or language mix is not supported, not validated, or not operationally practical with the offline models and hardware available.

For cloud multilingual workflows, Gemini, Claude, GPT-class, and other language-validated cloud models are strong candidates. For offline multilingual work, Qwen-family models, Gemma, and Mistral-family models each have distinct strengths.

11.3 Suggested model choices by language need

Language need	Suggested starting point
Primary language supported by an available offline model	Start with that offline model and validate on a representative sample.
Primary language not supported or not validated offline	Use a cloud model with strong support for that language, subject to provider and data-residency approval.
Broad mixed-language or unknown-language corpus	Gemini 3.1 Flash-Lite or Gemini 3 Flash (preview); consider Claude for premium analysis, or run a language-identification pass first.
Korean-heavy cloud matter	Gemini 3.1 Flash-Lite (either deployment) and Gemini 3 Flash (preview) both achieved perfect Tests 3 and 4; Claude 4.7 Opus achieved perfect Test 4 with one Test 3 error.

Korean-heavy offline matter	Qwen 3.6 27B Dense or Gemma 4 26B Think (both perfect or near-perfect on Korean Tests 3 and 4); Qwen 3.6 35B MoE as a faster sibling with a small accuracy concession; Mistral Small 3.2 24B as a strong alternative on modest hardware.
French-heavy offline matter	Mistral-family models, with validation.
Chinese/Japanese offline matter	Qwen-family models, with validation.
English-dominant high-accuracy offline matter	Mistral Small 3.2 24B or Llama 3.3 70B, depending on hardware and throughput.

11.4 Prompting multilingual matters

A single prompt written in the reviewer's working language can often classify emails in other languages, but testing is required. For some matters, especially where legal or cultural nuance is important, a translated prompt or jurisdiction-specific language may improve results.

For multilingual datasets:

- Include representative samples from each language.
- Review false positives and false negatives by language.
- Consider separate runs for major languages if the prompt behaves differently.
- Avoid assuming performance on one primary language predicts performance on another.

12. Quality Control, Reproducibility, and Defensibility

AI-assisted classification should be treated like any other technical evidence-processing method: document the process, validate the configuration, preserve the result, and review edge cases.

12.1 Defensible workflow pattern

```
Define issue
   ↓
Write or adapt prompt
   ↓
Select model and deployment path
   ↓
Run representative sample
   ↓
Review Classify labels or Filter passed/rejected records, plus Analyze accuracy
where applicable
   ↓
Revise prompt if needed
   ↓
Run production job
   ↓
Archive prompt, settings, model, provider, output, and validation notes
```

12.2 Review the right sets

At minimum:

- Review all Responsive results for obvious false positives.
- Review all INCONCLUSIVE results.
- Spot-check Unresponsive results.
- Pay special attention to uncommon languages, large attachments, and forwarded threads.

Design the validation sample to match the risk profile of the matter. For higher-stakes workflows, use stratified sampling across custodians, time periods, folders, languages, attachment types, and expected issue categories. For low-prevalence matters, include seeded known-positive or likely-positive examples where available, because a random sample may contain too few responsive items to test recall meaningfully.

Record the sampling rationale, sample size, acceptance criteria, categories reviewed, reviewer notes, false positives, false negatives, INCONCLUSIVE handling, named approver, and any prompt changes made after validation.

For sample validation, HTML output with the portable email viewer included is a practical review format. The browser viewer exposes AI classification results in the email list and supports tagging, making it easier to mark false positives, false negatives, INCONCLUSIVE items, and examples for prompt revision.

If viewer tags are used, export and archive the tag file before treating those tags as validation, QA, escalation, or review-decision records.

Where stakes are high, increase sampling and consider a second model for comparison.

12.3 Understand classification errors versus hallucination

For constrained Filter and Classify workflows, the main AI risk is usually a **classification error**, not a hallucination.

A classification error occurs when the model chooses the wrong label, often because the email is ambiguous, the prompt is unclear, categories overlap, or the model misunderstands context.

A hallucination would mean the model fabricated content not present in the source email. Aid4Mail's constrained Filter and Classify workflows reduce this risk by tying every output to a specific email and limiting the model to defined labels or structured task outputs.

Analyze tasks require additional review because their outputs are less constrained. Summaries, translations, extracted fields, issue notes, sentiment indicators, and risk explanations should be checked for omissions, unsupported statements, mistranslations, incorrect extraction, and statements not grounded in the source email.

For review purposes, treat AI output as an auditable decision or analysis associated with a source email, not as an independent fact.

12.4 Reproducibility

Some models and settings may not reproduce every classification exactly if rerun later. Reasons include provider-side model updates, internal reasoning behavior, nondeterministic outputs, and changes in infrastructure.

In the retained benchmark set, the main reproducibility distinction is between deterministic non-reasoning models and models that may vary between runs because of reasoning behavior, provider-side temperature handling, or both.

Affected retained models include:

- **Cloud:** Claude 4.7 Opus, OpenAI GPT-5.4.
- **Offline:** Qwen 3.6 27B Dense, Qwen 3.6 35B MoE, Gemma 4 26B Think.

If strict reproducibility is required, prefer deterministic retained alternatives such as:

- **Cloud:** Grok 4.2 Non-Reasoning, Gemini 3.1 Flash-Lite.
- **Offline:** Mistral Small 3.2 24B, Llama 3.3 70B, Ministral 3 14B.

Gemini 3 Flash (preview) is retained, but the supplied reproducibility guidance does not identify it as a strict reproducibility alternative. Treat it as uncharacterized for exact rerun reproducibility unless your own validation confirms repeatability.

Run-to-run variability is typically most relevant for ambiguous boundary emails. For defensibility, treat the exported output generated at the time of the production run as the controlling record.

Preserve:

- Exported classifications or analysis fields.
- Aid4Mail processing logs and AI Filter passed/rejected decision records, if AI Filter was used.
- Prompt text.
- Model name and provider.
- Provider platform and region.
- Date and time of run.
- Attachment settings.
- For offline runs: local model tag/name, quantization, model digest or file hash where available, local inference tool and version, endpoint URL, configured context length, and confirmation of the model loaded at run time.
- Any sample-validation notes.

If strict reproducibility is required, prefer models and settings known to support deterministic behavior, and preserve the exact output rather than relying on reruns.

12.5 Use INCONCLUSIVE as a safety boundary

INCONCLUSIVE is useful when the model identifies a potentially relevant signal but cannot make a reliable committed decision.

Do not ignore INCONCLUSIVE results. They should be routed to human review or a second-pass workflow.

If your workflow treats INCONCLUSIVE as Unresponsive, you are changing the risk profile and should re-evaluate recall accordingly.

12.6 Consider second-pass review for high-risk matters

For high-risk matters, consider:

- Running a second model on Responsive and INCONCLUSIVE results.
- Running a second model on a sample of Unresponsive results.
- Using a premium model only for the gray-zone subset.
- Comparing classifications across two model families.
- Reviewing all items where models disagree.

This can control cost while improving confidence.

---

13. Troubleshooting

Most AI workflow problems fall into a few categories: credentials, provider quota, context limits, local-server status, prompt design, or output configuration.

13.1 Authentication errors

Check:

- API key is valid.
- Provider account has available credit.
- Correct project, resource, or region is configured.
- Google Cloud service account file exists and has the required role.
- AWS IAM credentials are valid for the selected Bedrock region/model.
- Azure resource name and API key are correct for Microsoft Foundry.

13.2 Local model connection errors

If Aid4Mail cannot connect to Ollama or LM Studio:

- Confirm the local inference server is running.
- For Ollama, run `ollama serve` if needed.
- For LM Studio, start the server from the Developer tab.
- Confirm the endpoint URL and port.
- Confirm the selected model is downloaded and loaded.
- Confirm the context length in Aid4Mail matches the local tool configuration.

13.3 Rate limits and throttling

HTTP 429 usually means provider quota or rate limits have been reached.

Aid4Mail may retry automatically, but persistent throttling may require:

- Higher provider tier.

- Increased quota.
- Enterprise platform deployment.
- Provisioned throughput where available.
- Smaller batch sizes.
- A different model or provider.

Some providers also return quota-like errors when the account balance is depleted or the selected region has no active quota for that model.

13.4 Context window errors

If emails exceed the model's context window:

- Reduce attachment inclusion.
- Lower attachment text-size limits.
- Use deterministic scoping or exclusion filters first, without using issue keywords as an unvalidated sole gate into AI review.
- Shorten prompts where possible.
- For offline models, confirm the configured context length is not too small for the payload but not so large that it forces CPU offload.
- Use a model with a larger context window only when truncation remains material after the steps above.

For slow offline processing, reduce context length before changing models. Oversized local context settings are a common cause of poor throughput.

13.5 Unexpected output

If classifications or analysis fields do not appear as expected:

- Confirm the correct AI task is configured.
- Confirm the prompt was saved.
- Confirm the model is available.
- For Classify, check the folder structure template and category list.
- For Analyze, check the selected output fields and target format.
- Run a small sample and inspect the exported AI field, classification folder, provider error messages, and processing log.

13.6 Too many false positives

Tighten the prompt:

- Define Responsive more narrowly.
- Add explicit exclusions.
- Add a forwarded-content rule.
- Require direct participant-authored evidence.
- Tell the model to default to Unresponsive unless criteria are clearly met.
- Validate on known false-positive examples.

13.7 Too many missed items

Broaden or clarify the prompt:

- Add examples of indirect language.
- Add synonyms and behavioral indicators.
- Remove overly strict wording.
- Check whether attachment content is needed.
- Check whether the selected model handles the corpus language well.
- Review whether pre-filtering removed relevant items before AI processing.

13.8 Too many INCONCLUSIVE results

Clarify the decision threshold:

- Define when INCONCLUSIVE is allowed.
- Add category precedence rules.
- Separate broad multi-category prompts into focused binary passes.
- Consider a stronger model for gray-zone items.

13.9 Prompt library missing

If the prompt library does not appear after selecting **Open**, check whether Windows Controlled Folder Access blocked installation of the prompt files. Inspect the **AI Prompts** subfolder of the Aid4Mail program folder and restore or reinstall the prompt files if needed.

---

14. Glossary

Term	Meaning in this guide
AI model	The system that reads the email payload and returns a classification or analysis result.
Large language model / LLM	A type of AI model trained to process and generate language. In Aid4Mail, it is used for structured email tasks.
Prompt	The instruction telling the AI model what to decide or produce.
Token	A unit of text used by AI models for capacity and billing. Input tokens are the email/prompt sent to the model; output tokens are the model's response.
Context window	The maximum amount of text a model can consider in one request.
Precision	Of the emails flagged Responsive, the percentage that were actually responsive. High precision means fewer false positives.
Recall	Of all truly responsive emails, the percentage the model found. High recall means fewer missed emails.

F1	A combined effectiveness score balancing precision and recall. In the Aid4Mail benchmark, F1 is computed over emails where the model committed to Responsive or Unresponsive; INCONCLUSIVE and NO_RESULT items are tracked separately through Decision Rate and Automation Yield.
Automation Yield	The share of the full corpus the model resolved correctly and confidently without downstream review, computed over the full corpus rather than only committed decisions.
Decision rate	The share of emails classified as Responsive or Unresponsive rather than INCONCLUSIVE or error. A high F1 with a low decision rate can still leave meaningful work for reviewers.
INCONCLUSIVE	A result for ambiguous emails that should go to human review or a second-pass workflow.
Offline model	A model running on local hardware through tools such as Ollama or LM Studio.
VRAM	GPU memory. Larger or more capable offline models generally need more VRAM.
Quantization	A compressed model format that reduces memory use and often increases speed for offline models.
Q4-class	A shorthand for 4-bit quantized local model formats used to reduce VRAM requirements.
Q4_K_M	A common offline model quantization format recommended for several retained models.
KV cache	GPU memory reserved for the model's attention state during inference. Larger context lengths reserve more KV-cache memory and can reduce throughput or force CPU offload.
MoE	Mixture of Experts: a model architecture where only part of the full model is active for each token, even though the full model still affects memory requirements.
EDRM / Electronic Discovery Reference Model	a framework defining stages from information governance through production.
MIH / Message Identification Hash	An EDRM specification used in the legal and eDiscovery industries to easily identify and match duplicate email messages.
Endpoint	The API address Aid4Mail uses to connect to a cloud provider or local AI server.
Rate limit / quota	A provider-imposed limit on requests or tokens per minute/day/month.
Data residency	The geographic location where data is processed or stored.

Appendix A: First-Run Checklist

Use this checklist before the first AI run in a matter.

Matter setup

- [ ] Matter objective is defined.
- [ ] AI use case is appropriate for semantic classification or analysis.
- [ ] Deterministic filters have been applied first where practical.
- [ ] Issue-keyword filters were not used as the sole AI gate unless recall impact was validated and documented.
- [ ] Required languages are known or sampled.
- [ ] Attachment strategy is defined.
- [ ] HTML output with portable email viewer is selected for prompt-validation review, if useful.

Provider/model setup

- [ ] Provider or offline model is configured.
- [ ] Provider terms and costs are understood.
- [ ] Service-account files and API keys are stored securely.
- [ ] Region/data-residency requirements are satisfied.
- [ ] Model is suitable for the task and corpus language.
- [ ] Quotas are sufficient for the expected job size.

Prompt setup

- [ ] Prompt defines Responsive criteria.
- [ ] Prompt defines Unresponsive criteria.
- [ ] Prompt defines INCONCLUSIVE criteria.
- [ ] Prompt handles forwarded or quoted content.
- [ ] Category list is mutually exclusive, if using classification.
- [ ] Prompt has been verified.

Sample validation

- [ ] Representative sample selected.
- [ ] Sample exported to HTML with portable email viewer included, if browser-based review and tagging are desired.
- [ ] Viewer tags exported and archived, if tagging was used.
- [ ] Sample includes known or likely responsive examples, not only random items.
- [ ] Production acceptance criteria and approver are defined.
- [ ] Low-prevalence risk considered and documented.
- [ ] Unresponsive sampling strategy documented.
- [ ] Responsive results reviewed.
- [ ] INCONCLUSIVE results reviewed.
- [ ] Unresponsive results spot-checked.
- [ ] Prompt revised if needed.

- [ ] Final prompt saved.

Production

- [ ] Incremental processing enabled where appropriate; if pre-acquisition filtering is used, the incremental-processing limitation is documented.
- [ ] Output format includes required AI fields.
- [ ] Run date/time recorded.
- [ ] Prompt/model/provider/settings recorded.
- [ ] Final output archived.
- [ ] HTML viewer tags or exported tag files preserved, if used during validation or review.
- [ ] Aid4Mail processing logs and AI Filter passed/rejected decision records preserved, if AI Filter was used.

Appendix B: Defensibility Checklist

Preserve the following for a production AI-assisted workflow.

Item	Record
Matter name / ID	
Corpus source	
Collection method	
Pre-acquisition filters	
Issue-keyword gate validation, if used	
Post-acquisition filters	
AI task used	Filter / Classify / Analyze
Provider	
Model	
Cloud region or offline host	
Prompt text	
Prompt version	
Category list	
Attachment inclusion setting	
Attachment size limit	
Context length, if offline	

Offline model tag/name, quantization, and digest/file hash if available	
Local inference tool/version and endpoint, if offline	
Confirmation of model loaded at run time, if offline	
Sample size	
Sampling design and rationale	
Known-positive or likely-positive seed examples used, if any	
Sample-validation notes	
Production acceptance criteria	
Reviewer / approver	
Production run date/time	
Export format	
Aid4Mail processing logs and AI Filter decision record, including passed and rejected items, if Filter was used	
HTML portable viewer used for sample review	Yes / No
Exported viewer tag file, if viewer tags were used	
Classification or analysis output archived	Yes / No
INCONCLUSIVE review process	
Notes on errors or throttling	
Raw AI output or provider response log, if available	
Second-model comparison notes, if used	

Appendix C: Benchmark Summary

The Aid4Mail AI Classification Benchmark evaluated retained cloud and offline models on realistic email-classification tasks relevant to digital forensics and eDiscovery.

C.1 Core accuracy result

On the primary 2,000-email binary responsiveness test:

Result	Meaning
Every retained model exceeded 94.5% F1	All retained models performed strongly on the core task.

Top five F1 scores ranged from 98.8% to 99.6%	Best models clustered tightly at very high effectiveness, with a top-five spread of about 0.8 percentage points.
Every retained model reached at least 99% recall	Missed evidence was rare in the benchmark operating point.
Offline and cloud models both appeared among top performers	Deployment mode did not determine accuracy.

F1 and Automation Yield measure different things. F1 measures the quality of committed Responsive/Unresponsive decisions. INCONCLUSIVE and NO_RESULT items are not counted as F1 errors in the benchmark; they are tracked separately through Decision Rate and Automation Yield. A model can have a high F1 while still leaving more items for human review if it abstains often.

C.2 Operational result

Finding	Meaning
Fastest tested offline model: ~879,000 emails/weekend, Test 5 production-payload basis	Offline AI can be extremely fast on suitable tasks and hardware.
Balanced offline model: ~614,000 emails/weekend	High accuracy and high throughput are possible locally.
Fastest measured cloud deployment: ~384,000 emails/weekend, Test 5 production-payload basis	Gemini 3.1 Flash-Lite (Agent Platform) was the fastest measured cloud deployment. The AI Studio deployment measured roughly ~289,000 emails/weekend and remains the lowest-cost cloud path.
Low-cost cloud estimate: roughly $42 per 100,000 emails, Test 5 production-payload basis	Cloud classification can be inexpensive compared with traditional review workflows; the Agent Platform deployment was roughly $43.
Offline marginal provider cost: $0	Hardware and IT costs still apply, but there is no token bill.
Top Automation Yield tie showed a speed contrast	Llama 3.3 70B offline and Grok 4.2 Non-Reasoning cloud both reached 99.85% AY on Test 1; the cloud run was about 9.6× faster on the reference comparison.

Throughput figures are not all based on the same test. Test 5 has eight measured deployment rows across seven model families: Ministral 3 14B, Mistral Small 3.2 24B, both Gemini 3.1 Flash-Lite deployments, Gemini 3 Flash (preview), OpenAI GPT-5.4, Gemma 4 26B Think, and Llama 3.3 70B. Grok 4.2 Non-Reasoning, Claude 4.7 Opus low effort, Qwen 3.6 27B Dense, and Qwen 3.6 35B MoE were added after Test 5 and currently use Test 1-derived throughput estimates.

C.3 Multilingual result

The benchmark directly tested English, French, Spanish, and Korean content. All retained models supported multilingual classification, but performance varied by model and language.

Several cloud and offline models performed strongly, including on Korean tests. For international matters, choose a model validated for the corpus's primary language or languages.

C.4 Caveats

Benchmark results are not guarantees. Real-matter performance depends on:

- Corpus composition.
- Responsive prevalence.
- Prompt clarity.
- Attachment strategy.
- Language mix.
- Model choice.
- Provider region and quota.
- Sampling and QC practices.

Use the benchmark to select a starting model, not to replace matter-specific validation.

Appendix D: Companion Documents

Use these companion documents for material that changes more often or requires more technical depth than this workflow guide.

D.1 Aid4Mail AI Provider and Model Configuration Guide

Use **Aid4Mail_AI_Provider_and_Model_Configuration_Guide** for:

- API keys.
- Google Vertex AI / Gemini Enterprise Agent Platform service-account setup.
- Amazon Bedrock IAM configuration.
- Microsoft Foundry resource names.
- Ollama and LM Studio setup.
- Local endpoints.
- Provider configuration files.
- Service-account security.
- Model configuration and placeholder mechanics.

D.2 Aid4Mail AI Provider and Model Selection Guide

Use **Aid4Mail_AI_Provider_and_Model_Selection_Guide** for:

- Model-by-model recommendations.
- Current provider pricing.
- Current regional availability.
- Context windows.
- Throughput estimates.

- Hardware requirements.
- Multilingual notes.
- Reproducibility notes.

This reference should be updated more frequently than the main workflow guide because provider pricing, model availability, model context windows, and regional support can change.

D.3 Aid4Mail AI Classification Benchmark Report

Use **AI_Classification_Benchmark_Report** for:

- Full methodology.
- Test design.
- Corpus details.
- Metrics.
- Per-model results.
- Excluded models.
- Production Pilot details.
- Benchmark limitations.

D.4 Prompt Library / Prompt Cookbook

Use **Prompt_Cookbook_Best_Practices_Guide** for:

- Digital Forensics prompt examples.
- eDiscovery prompt examples.
- FOIA/Public Records prompt examples.
- Prompt validation patterns.
- Common category-taxonomy designs.
- Examples of overbroad and corrected prompts.

Conclusion

Aid4Mail AI should be used the same way other serious forensic and eDiscovery technologies are used: deliberately, with validation, documentation, and review. The advantage is not that AI removes professional judgment. The advantage is that it lets practitioners apply that judgment at scale, using plain-language criteria, while preserving outputs that can be reviewed, exported, and defended.

Date of publication: May 15, 2026.