Jylo

AI Model Card

A guide to the AI models available on the Jylo platform

Version 1.0

Introduction

Jylo provides access to a curated selection of Artificial Intelligence (AI) models from leading providers. Each model has different strengths, and choosing the right one for your task can improve both the quality and efficiency of your analysis.

This guide explains the models available on the platform, what each one is best suited to, and what to keep in mind when reviewing AI-generated outputs. It is designed to support your organisation's transparency, governance and compliance obligations—including the Algorithmic Transparency Recording Standard (ATRS).

> **Tip**
> Not sure which model to choose? For everyday tasks with well-defined instructions, start with Claude Haiku 4.5 or GPT-4.1—both are fast, reliable and cost-effective. For tasks that are less well defined and require reasoning or intuition, step up to Claude Sonnet 4.6 or GPT-5. For ambitious, complex work where you need the deepest analysis, choose Claude Opus 4.6.

How Jylo Uses AI Models

Jylo does not build or train its own AI models. Instead, it provides a secure, governed workspace in which commercially available models are accessed via API to perform document analysis tasks that you define.

You interact with AI models in two ways on Jylo:

- **Assistant:** A conversational chat interface where you can ask questions about your documents, compare files and conduct ad-hoc analysis.
- **Playbooks and Flows:** Structured, repeatable workflows where your prompts are applied to one or many documents individually, with results subject to human verification.

In both cases, you choose which model to use, what questions to ask and whether to accept the AI's output. Your data is never used to train AI models, and all interactions remain within your organisation's secure environment.

Available Models

Jylo offers several models across three providers. Your administrator controls which models are available in your region. Models convert text into tokens for processing; each model has a token limit. 100 thousand tokens are equivalent to roughly 75 thousand words.

GPT-5

Provider	OpenAI (hosted on Microsoft Azure)

Best for	Tasks requiring reasoning and intuition—regulatory analysis, nuanced legal interpretation, multi-step logic where instructions may be less well defined
Strengths	Near-frontier intelligence. Strong reasoning and multimodal capabilities. Excellent for complex, multi-step analysis.
Limitations	Significantly slower than other models. Affordable credit consumption, especially for a reasoning model.
Context window	400,000 tokens
Knowledge cutoff	September 2024

GPT-4.1

Provider	OpenAI (hosted on Microsoft Azure)
Best for	Everyday tasks with well-defined instructions—summarisation, question-answering, data extraction and lengthy document review.
Strengths	One of the fastest models on the platform. Largest context window (1M tokens)—ideal for very long documents. Strong all-round performer.
Limitations	Less suited to tasks requiring deep multi-step reasoning compared with GPT-5. May require tailored prompting for highly specialist terminology.
Context window	1,000,000 tokens
Knowledge cutoff	June 2024

> **Tip**
> GPT-4.1 and Claude Haiku 4.5 are recommended for everyday tasks with clear, well-defined instructions. GPT-4.1's advantage is its enormous context window (1M tokens), making it the best choice when working with very long documents.

o3-mini

Provider	OpenAI (hosted on Microsoft Azure)
Best for	Fast, affordable reasoning tasks—structured analysis, logic-driven prompts and compliance checks where cost matters
Strengths	Reasoning model that "thinks through" problems step by step. Moderate end-to-end speed (approx. 11 seconds including thinking time). Cost-effective.
Limitations	Lower overall intelligence than GPT-5 or Claude Opus 4.6. Not recommended for the most complex, high-stakes analysis.
Context window	200,000 tokens
Knowledge cutoff	October 2023

Claude Opus 4.6

Provider	Anthropic (hosted on Amazon Web Services)
Best for	Ambitious, complex work requiring the deepest analysis—multi-document reasoning, high-stakes compliance, nuanced interpretation where maximum intelligence matters
Strengths	Highest intelligence of any model on the platform. Strong reasoning and low hallucination rates. Well suited to detailed legal and regulatory work.
Limitations	Highest credit consumption among Claude models. Slower response times for very large inputs.
Context window	200,000 tokens
Knowledge cutoff	Early 2025

Claude Sonnet 4.6

Provider	Anthropic (hosted on Amazon Web Services)
Best for	Tasks requiring reasoning and intuition at a moderate cost—contract review, regulatory analysis and less well-defined questions where the model needs to interpret intent
Strengths	Near-frontier intelligence at moderate cost. Strong reasoning capabilities without the premium cost of Opus.
Limitations	Does not match Opus on the most complex tasks. Slower than some OpenAI models.
Context window	200,000 tokens
Knowledge cutoff	Early 2025

Claude Haiku 4.5

Provider	Anthropic (hosted on Amazon Web Services)
Best for	Everyday tasks with well-defined instructions—document triage, classification, structured extraction, batch processing and cost-sensitive workflows
Strengths	Fastest model on the platform. Strong intelligence for its cost—outperforms GPT-4.1 and o3-mini on benchmark intelligence. Ideal when speed, volume and value matter.
Limitations	Smaller context window than GPT-4.1. Less suited to highly complex multi-step reasoning compared with Opus or Sonnet.
Context window	200,000 tokens
Knowledge cutoff	April 2024

LLama 3.3

Provider	Meta (hosted on Microsoft Azure)
Best for	Budget-friendly, high-volume tasks—simple classification, rapid triage and straightforward extraction where cost is the priority
Strengths	Open-source model. Fast and cost-effective. Good for large batches of straightforward tasks.
Limitations	Lower intelligence than proprietary models. Not suited to complex legal, regulatory or multi-step analysis. Best paired with thorough human review.
Context window	128,000 tokens
Knowledge cutoff	December 2023

Quick Comparison

Use this table to compare models at a glance:

Model	Provider	Intelligence	Speed (E2E)	Cost	Best For
GPT-5	OpenAI	★★★★	★	£½	Reasoning and intuition
GPT-4.1	OpenAI	★★★	★★★★★	£½	Everyday, well-defined tasks
o3-mini	OpenAI	★★★	★★★	£	Fast, affordable reasoning
Claude Opus 4.6	Anthropic	★★★★★	★★★	£££	Ambitious, complex work
Claude Sonnet 4.6	Anthropic	★★★★	★★★	££	Reasoning and intuition
Claude Haiku 4.5	Anthropic	★★★½	★★★★★	£	Everyday, well-defined tasks
LLama 3.3	Meta	★★	★★★★★	½	Budget high-volume tasks

What to Keep in Mind

All AI models have limitations. Understanding these helps you get the best results from Jylo and ensures your organisation uses AI responsibly.

Bias

AI models are trained on large bodies of text from the internet and published sources. This training data may contain historical, cultural or demographic biases that can surface in

outputs. Models may also default to US-centric legal or regulatory assumptions—if you work within UK or other jurisdictions, frame your prompts accordingly and verify outputs against local requirements.

Hallucination

Models can sometimes generate responses that sound authoritative but contain factual errors. This is why Jylo's verification workflow exists—every AI output in a Flow can be approved, modified or rejected by a human before it is treated as final.

Prompt Sensitivity

Small changes in how you phrase a question can produce different results. Jylo's Playbook system helps standardise your prompts so that the same methodology is applied consistently across documents.

> **Important**
> AI outputs on Jylo should always be treated as a starting point for human review, not as a final determination. Use the verification workflow to approve, modify or reject every output before relying on it.

How to Select a Model

In the Assistant

1. Click the **model dropdown menu** in the top-left corner of the Assistant interface.
2. Select your preferred model from the list.
3. The Assistant will use your selected model for all subsequent messages until you change it.

In Playbooks

Each prompt within a Playbook can be configured to use a specific model. When creating or editing a prompt, select the desired model from the model dropdown in the prompt configuration form.

Administrator Controls

Administrators control which models are available across the organisation. To configure model availability:

4. Navigate to **Admin > Settings > Models**.
5. Select the desired region tab.
6. Check or uncheck models to enable or restrict them.

Data Handling and Security

- **Your data is private:** Jylo never shares your data with third parties. Your data is never used to train AI models.
- **Infrastructure:** OpenAI and LLama models are hosted on Microsoft Azure. Anthropic Claude models are hosted on Amazon Web Services.
- **Access control:** Role-based permissions govern who can access Projects, Playbooks and model configurations.

For further information on Jylo's security posture and certifications, visit trust.jylo.ai.

Next Steps

- **Try different models:** Experiment with different models in the Assistant to see which one works best for your documents and questions.
- **Access our Guides:** For user guides and video demonstrations visit our support.jylo.ai.
- **Contact support:** For any questions, reach out to support@jylo.ai.

© 2026 Jylo. All rights reserved.