ORACLE

How Oracle Selects, Manages, and Bills for AI Models

Transparency on model choices, defaults, and consumption in Oracle Fusion Cloud Applications

This paper applies to Oracle Fusion Applications Release 26C and later.
October 2026, Version 1.3
Copyright © 2026, Oracle and/or its affiliates
Public

Executive summary

Oracle Fusion AI gives your organization a powerful set of AI models, selected and continuously evaluated by Oracle so that the agents you rely on deliver strong enterprise outcomes out of the box. This guide explains how Oracle chooses those models, how and when they're updated, the choices you control, and how AI consumption is measured and billed so you can better understand and plan for it.

Two principles sit underneath everything that follows. First, you can control the trade-off between quality and consumption, and for most use cases, you can compare models and switch at any time. Second, Oracle-built, seeded agents and execution of general action types (e.g., Q&A, summarization, and approvals) that runs on the basic model hosted on Oracle Cloud Infrastructure (OCI) is available to you at no incremental consumption cost.

How Oracle thinks about AI models

A large language model (LLM) is the reasoning engine behind an agent. The same agent—for example, a journal-entry reconciliation agent or an invoice-matching agent—can run on different underlying models, and the choice of model affects the result's quality (accuracy), its speed, and the amount of AI consumption it draws.

Oracle gives you a good starting point by curating a set of supported models, vetting each one for enterprise use, validating quality and performance for specific use cases, and assigning sensible defaults so that agents work well from the moment you turn them on. But you retain the ability to change those defaults when you want more control.

Oracle's model philosophy

Oracle's approach to model selection is guided by the following five principles.

1. Quality first by default

Oracle-built agents ship configured to help deliver accurate, reliable results for enterprise work out of the box. For many agents that means a higher-capability model is selected as the default, because it is designed to produce strong outcomes with less tuning. You can choose between frontier models (highly capable, proprietary models), balanced models (that optimize price and performance) or basic models (smaller, faster, and cheaper options).

Model class	What it's for	Example tasks	Consumption
Frontier	Maximum capability on hard problems	Deep research, long-running multi-step analysis, complex reasoning across large context	Highest
Balanced	Price-performance optimized; the default for most enterprise work	Drafting, summarization, classification, most agent workflows	Balanced
Basic	Simple, high-volume tasks at low consumption	Routine extraction, straightforward Q&A, high-volume lookups	Lowest

2. Curated, not unlimited

Rather than exposing every model on the market, Oracle maintains a vetted set of supported models and keeps them current. Each model is evaluated for performance, quality, safety, and fit before it is offered. Oracle prioritizes models that support multiple languages, and our current Oracle-provided models are multi-lingual, offering support for more than 90 languages.

Copyright © 2026, Oracle and/or its affiliates / Public

3. Choice where it matters

You can select a lighter-weight model for an agent, compare it against the default, and decide which model is right for each use case. Oracle also supports bringing your own model through supported providers (bring-your-own LLM, or BYO LLM). You can adjust the model used at different levels of the workflow (e.g., at the agent level, or the node level) to right-size your spend and balance cost with performance, depending on the type of task being done.

4. Feature-rich where enterprise work demands it

Oracle's model strategy goes beyond text generation. Supported models are selected to power a broad range of enterprise AI use cases, including multimodal understanding, document and image analysis, structured reasoning, code generation, tool use, retrieval-augmented generation, and agentic workflows. As model capabilities evolve, Oracle selects and integrates the features that matter most for business applications, enabling agents to handle richer inputs, more complex tasks, and more complete end-to-end workflows.

5. A low-cost path is available

Oracle-built, seeded agents are included with your subscription. What they consume depends on the model they use. With Oracle-provided, OCI-hosted models, execution of general action types (e.g., Q&A, summarization, and approvals) does not draw down your AI Unit allocation. With balanced models, seeded agents consume AI Units like any other agent. You can see which model an agent uses, and change it, at any time.

Plain language on default models

Some Oracle-built agents default to a higher-capability model because it produces the best result without configuration. That default optimizes quality, not consumption. If you would prefer to minimize consumption, you can test the agent on a lighter-weight model and switch to it. The agent keeps working, and on OCI-hosted models that execution does not draw down your allocation.

Model lifecycle and deprecation

Models are introduced, improved, and eventually retired. Oracle manages that lifecycle openly, so you can see a model's status, how long you have, and what to move to.

- **Deprecation is announced with a removal date.** When an Oracle-provided model is scheduled for retirement, Oracle announces a set removal date in advance and makes a replacement model available at the same time.
- **You have up to six months to migrate.** You may move at any point in that six-month window, and you are expected to complete migration before the removal date.
- **Status is visible in the product.** Deprecated models are labeled in AI Agent Studio with their removal date and shown alongside candidate replacement models, so you can see a model's standing before you assign it.
- **A safety net exists, but it isn't a substitute for migrating.** If a removed model is still referenced at runtime, the platform falls back to the best available model in the same class, so your agent does not fail outright. This does not guarantee equivalent output or performance, and every fallback event is logged. Test your agents against the replacement model rather than relying on fallback.

BYO LLMs follow your provider's lifecycle, not Oracle's. When a provider retires a model you have connected, you manage that transition under your own governance, and Oracle surfaces the model's status in AI Agent Studio so you can see where it stands.

Copyright © 2026, Oracle and/or its affiliates / Public

Model cards and what they show you

In AI Agent Studio, every available model is presented as a card. As of 26C, those cards show you key characteristics before you assign it to an agent. The same transparency you get on consumption now extends to the models themselves.

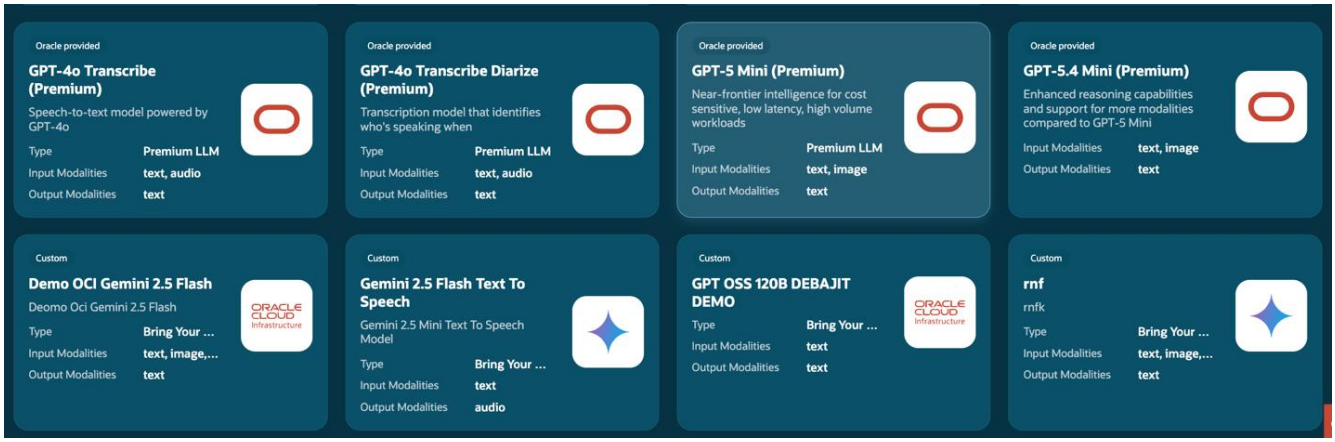


Figure 1. Model cards in AI Agent Studio

Each card tells you four things at a glance:

- **Who provides and maintains it.** Cards are labeled “Oracle provided” or “Custom.” Oracle-provided models are curated, vetted, and version-managed by Oracle on your behalf. Custom models are ones you bring through a supported provider (BYO LLM) and manage under your own governance and credentials.
- **Capability tier.** Higher-capability models are labeled “Balanced” both in the model name and in the Type field so the tier is clearly identified rather than implied. “Balanced” describes the model’s capability, not a separate charge; what you consume is always measured in AI Units, and execution of general action types (e.g., Q&A, summarization, approvals) on Oracle-provided, OCI-hosted models does not draw down your allocation.
- **What the model is good for.** A plain-language description from the model provider and the admin who configures it states each model’s intended use. For example, a model suited to “cost sensitive, low latency, high volume workloads,” or a transcription model that “identifies who’s speaking when.” This helps you match the model to the job rather than guess from the name.
- **What the model can handle.** Each card lists the model’s input and output modalities (text, audio, image), so you can see whether a model fits your use case before assigning it.

The cards turn model selection into an informed choice: you can see provenance, capability, purpose, and fit side by side, then use the controls explained below to assign, compare, or switch models as your needs change.

Copyright © 2026, Oracle and/or its affiliates / Public

ORACLE

The choices you control

For Oracle-built agents, your administrators have a consistent set of controls in AI Agent Studio. These choices are not permanent. You can revisit them as your needs change.

Your choice	What it means	Where you do it
Run on the default model	Use the model Oracle selected for best out-of-box quality. Recommended starting point.	AI Agent Studio (no action needed)
Switch to a different model	Test an agent's output on different models side by side, then choose the model that fits the use case or lowers your cost. (E.g., reduce consumption by staying on an OCI-hosted model when the default's extra capability isn't needed.) Connect a supported provider model, including a fine-tuned or region-specific model, under your own governance.	AI Agent Studio model comparison and configuration
Move ahead of deprecation	See whether a model is current or deprecated. If deprecated, see what replaces it, then test the successor and switch on your own timing. Oracle-provided models are transitioned for you.	AI Agent Studio model cards and comparison

Basic, OCI-hosted models are ideal for developing, testing, piloting and running simple use cases at modest volumes. These include embedded generative AI use cases such as text summarization and basic Q&A with simple, small text only documents. For best quality and performance (latency), and for higher volumes and rates of usage, we recommend leveraging the balanced, vendor-hosted models (from Open AI and/or Google), which are provided through AI Agent Studio.

Bringing your own model

You can run agents on a specific model from a supported provider. Supported providers currently include Azure, Google, OCI, OpenAI, and VertexAI. The model configuration page in AI Agent Studio shows every model enabled for the provider you select.

If the model you need is not yet on the list, you can request it through Oracle Support by opening a service request that includes the provider name, the requested model and version, your business use case, and your required timeline. (Keep in mind that you can only request models from supported providers.) Oracle reviews each request and provides an availability estimate. Simpler configuration updates are typically available in about two weeks; requests that require code changes typically take about four weeks. For all BYO LLMs, you manage your own third-party provider credentials and any usage costs those providers charge.

How and when Oracle updates models

The AI landscape moves quickly. Oracle continuously vets new and updated models and manages versions on your behalf, so the agents you depend on benefit from improvements without you having to track the market or re-engineer your configurations.

Oracle's model management includes:

- **Continuous vetting.** New and updated models are evaluated for quality, safety, and enterprise fit before they become available.

Copyright © 2026, Oracle and/or its affiliates / Public

- **Version management.** Oracle manages model versions so agents stay current. Changes to curated defaults (including deprecating models) are communicated through standard release channels.
- **Use-case optimized selection.** Models are matched to the work an agent does, rather than applying a single model to every task.
- **Governance controls across providers.** When you use multiple models, consistent authentication, access control, and auditability apply across providers.

How AI consumption is measured and billed

Oracle bills for the execution of enterprise work, not for AI components. Consumption is measured in AI Units, and what you pay reflects the work an agent performs on your behalf.

What is and isn't consumed

The following table outlines how different activities use (or don't use) AI Units.

Activity	Does it draw down AI Units?
Oracle-built, seeded agents	No; included at no cost when powered by the basic LLM and when taking general actions (e.g., Q&A, summarization, and approvals)
Execution on Oracle-provided, OCI-hosted models	No; general actions do not decrement your allocation
Agents and Fusion Agentic Applications running on metered models	Yes; measured in AI Units as work is executed

Because Oracle-provided, OCI-hosted model execution does not draw down your allocation, the lowest-consumption path and a high-quality path often coincide. When an agent does run on a metered model, consumption is reported in outcome terms so you can evaluate the trade-off between spend and performance.

Seeing your consumption in Release 26C and later

As of Release 26C, your AI Unit consumption per pod or instance is shown directly in AI Agent Studio, and Fusion SaaS usage, including your remaining balance, appears in the Cloud Console. You can see exactly what is being consumed, by which agents, and how much of your allocation remains.

Visibility is the heart of this release. You can see what is being consumed and what remains in two places:

- **AI Agent Studio:** AI Unit consumption is displayed per customer, so you can see usage as it happens and attribute it to specific agents and apps.
- **Cloud Console:** Fusion SaaS usage reporting displays your AI usage totals, including your remaining balance.

AI Units begin decrementing in the month you receive 26C. For customers who purchased the new subscription beginning in May 2026, billing for AI Units starts in the month you receive 26C and is tied to your use of agents and Fusion Agentic Applications. In the interim, AI Units used in development, staging, and production are not charged for these early customers.

Copyright © 2026, Oracle and/or its affiliates / Public

If you are on the earlier FY26 subscription

Customers who purchased the earlier FY26 subscription, which used a different metering approach and custom agents, can continue with it, but Oracle encourages moving to the current AI Unit subscription because it aligns billing to outcomes and gives you the new consumption visibility described above. For customers who remain on the previous subscription, Oracle monitors usage that runs well beyond what is included and works with those customers individually. The preferred path in those cases is moving to the current AI Unit products; where needed, the prior add-on option remains available.

Line of sight into what's coming

Oracle is expanding the ways you build and run agents, and broadening model choice and visibility over time. For transparency, it's important to understand what's coming:

- **Builder experience.** Building agents is becoming progressively easier starting with no-code, then extending to pro-code and additional no-code paths so that more of your team can assemble and tune agents without deep technical work.
- **Expanding model choice.** Oracle continues to add supported providers and models, including speech-to-text and text-to-speech capabilities, and to streamline how you request models that aren't yet listed.
- **Deeper visibility.** Consumption and model insight in AI Agent Studio and the Cloud Console are expected to grow, building on the per-customer visibility introduced in 26C.

Oracle's approach

Oracle's goal is to make AI model selection predictable, transparent, and aligned to your enterprise needs. The following principles outline Oracle's approach:

- Oracle selects and maintains models to help your agents deliver strong enterprise outcomes by default.
- Model choices are visible and reversible. You can compare and adjust models based on the quality/consumption trade-off.
- A capable, low cost path is available through seeded agents and Oracle-provided, OCI-hosted models.
- AI consumption and remaining allocation are visible in AI Agent Studio and the Cloud Console.
- Customers are notified when models change and informed about the future direction of Fusion AI.

As model choice and capabilities continue to expand, these principles are intended to guide how Oracle selects, manages, and supports AI models in Fusion Applications.

Connect with us

Call +**1.800.ORACLE1** or visit **oracle.com**. Outside North America, find your local office at: **oracle.com/contact**.

blogs.oracle.com facebook.com/oracle twitter.com/oracle

Copyright © 2026, Oracle and/or its affiliates. This document is provided for information purposes only, and the contents hereof are subject to change without notice. This document is not warranted to be error-free, nor subject to any other warranties or conditions, whether expressed orally or implied in law, including implied warranties and conditions of merchantability or fitness for a particular purpose. We specifically disclaim any liability with respect to this document, and no contractual obligations are formed either directly or indirectly by this document. This document may not be reproduced or transmitted in any form or by any means, electronic or mechanical, for any purpose, without our prior written permission.

Oracle, Java, MySQL, and NetSuite are registered trademarks of Oracle and/or its affiliates. Other names may be trademarks of their respective owners.