Yandex AI Studio

- Начало работы с Model Gallery
- Концепции
  - О сервисе Yandex AI Studio
  - Model Gallery
    - Обзор
    - Модели базового инстанса
    - Модели выделенного инстанса
    - Пакетная обработка данных
    - Вызов функций
    - Режим рассуждений
    - Форматирование ответов моделей
    - Классификаторы
    - Эмбеддинги
    - Датасеты
    - Дообучение
    - Токены
  - Agent Atelier
  - AI Search
  - MCP Hub
  - Yandex Workflows
  - Квоты и лимиты
  - Термины и определения

Модели базового инстанса

Статья создана Yandex Cloud Обновлена 24 ноября 2025 г.

Сервис Yandex AI Studio предоставляет доступ к большим генеративным моделям, разработанных разными компаниями. Если стандартных моделей вам недостаточно, вы можете дообучить некоторые модели, чтобы они точнее отвечали на ваши запросы. Все роли, необходимые для работы с моделями, перечислены в разделе Управление доступом в Yandex AI Studio.

В базовом инстансе ресурсы модели доступны всем пользователям Yandex Cloud и делятся между ними, поэтому при большой нагрузке время работы моделей может увеличиваться. При этом другие пользователи гарантированно не могут получить доступ к контексту ваших переписок с моделью: даже при включенном режиме логирования запросы хранятся в обезличенном виде, а потенциально чувствительная информация маскируется. Однако если вы обрабатываете конфиденциальную информацию с помощью моделей, рекомендуем отключать логирование данных.

Для моделей базового инстанса действуют правила обновления, описанные в разделе Жизненный цикл модели. При обновлении моделей поколения, доступные в разных ветках (сегменты `/latest`, `/rc` и `/deprecated`), могут меняться. Модифицированные модели делят квоты на использование со своими базовыми моделями.

Модель и URI	Контекст	Доступные API
Alice AI LLM `gpt://<идентификатор_каталога>/aliceai-llm`	32 768	API генерации текста, OpenAI-совместимые API
YandexGPT Pro 5.1 `gpt://<идентификатор_каталога>/yandexgpt/rc`	32 768	API генерации текста, OpenAI-совместимые API
YandexGPT Pro 5 `gpt://<идентификатор_каталога>/yandexgpt/latest`	32 768	API генерации текста, OpenAI-совместимые API
YandexGPT Lite 5 `gpt://<идентификатор_каталога>/yandexgpt-lite`	32 768	API генерации текста, OpenAI-совместимые API
Qwen3 235B `gpt://<идентификатор_каталога>/qwen3-235b-a22b-fp8/latest`	262 144	OpenAI-совместимые API
gpt-oss-120b `gpt://<идентификатор_каталога>/gpt-oss-120b/latest`	131 072	OpenAI-совместимые API
gpt-oss-20b `gpt://<идентификатор_каталога>/gpt-oss-20b/latest`	131 072	OpenAI-совместимые API
Дообученная YandexGPT Lite `gpt://<идентификатор_каталога>/yandexgpt-lite/latest@<суффикс>`	32 768	API генерации текста, OpenAI-совместимые API
Gemma 3 27B `gpt://<идентификатор_каталога>/gemma-3-27b-it/latest` Условия использования Gemma	131 072	OpenAI-совместимые API
YandexART `art://<идентификатор_каталога>/yandex-art/latest`	500 символов	API генерации изображений

Модель Gemma 3 27B работает с изображениями в кодировке Base64. Модель может обрабатывать изображения с любым соотношением сторон благодаря адаптивному алгоритму, который масштабирует изображения до 896 пикселей по большей стороне, сохраняя важные визуальные детали. Каждое изображение использует 256 токенов контекста.

Жизненный цикл модели

Каждая модель имеет набор характеристик жизненного цикла: название модели, ветка и дата публикации. Эти характеристики позволяют однозначно определить версию модели. Обновление моделей происходит по определенным ниже правилам, чтобы вы могли адаптировать свои решения под новую версию, если это будет необходимо.

Существует три ветки модели (от более старой к новой): `Deprecated`, `Latest`, `Release Candidate` (`RC`). Для каждой из этих веток действует SLA сервиса.

Ветка `RC` обновляется по мере готовности новой модели и может измениться в любой момент. Когда модель в ветке `RC` будет готова к общему использованию, в истории изменений и сообществе пользователей в Telegram появится уведомление о предстоящем релизе.

Через месяц после объявления версия `RC` становится `Latest`, а `Latest` переносится в `Deprecated`. Поддержка версии `Deprecated` осуществляется в течение следующего месяца, после чего модели в ветках `Deprecated` и `Latest` будут идентичны.

Обращение к моделям

Вы можете обращаться к моделям генерации текста разных версий несколькими способами.

SDK **API**

При работе с моделями генерации текста через Yandex Cloud ML SDK используйте один из следующих форматов:

- **Название модели**, передается в виде строки. Доступны только версии `Latest`.

```
# Генерация текста
model = (
  sdk.models.completions("yandexgpt")
)

# Генерация изображений
model = (
  sdk.models.image_generation("yandex-art")
)
```

- **Название и версия модели**, передаются в виде строк в полях `model_name` и `model_version` соответственно.

```
# Генерация текста
model = (
  sdk.models.completions(model_name="yandexgpt-lite", model_version="rc")
)

# Генерация изображений
model = (
  sdk.models.image_generation(model_name="yandex-art", model_version="latest")
)
```

В приведенном примере явно заданы модели `YandexGPT Lite` версии `Release Candidate` и `YandexART` версии `Latest`.

- **URI модели**, передается в виде строки, содержащей полный URI нужной версии модели. Также используйте этот способ для обращения к дообученным моделям.

```
# Генерация текста
model = (
  sdk.models.completions("gpt://b1gt6g8ht345********/yandexgpt/deprecated")
)

# Генерация изображений
model = (
  sdk.models.image_generation("art://b1gt6g8ht345********/yandex-art/latest")
)
```

В приведенном примере явно заданы модели `YandexGPT Pro` версии `Deprecated` и `YandexART` версии `Latest`.

См. также

- Отправить запрос в синхронном режиме
- Отправить асинхронный запрос
- Сгенерировать изображение с помощью YandexART
- Запустить модель в пакетном режиме