Backyard.ai для самых маленьких

Если вы когда-либо хотели себе чатбота, который не будет зависеть от качества интернета, рабочести серверов вашего любимого сайта, не будет плеваться от слова "бибизьга" и не будет пытаться вытянуть из вас последние шекели, то вы, наверное думали о том, как бы так попрощей да побыстрей накатить себе на комп личную нейросеточку, но натыкались только на гайд, как устанавливать SillyTavern, для которго нужно накатить Git, питон, литр водки и девственной крови. И было неясно, разберётесь вы с этим, да и стоит ли игра свеч. И нужно ли оно вам вообще.

Когда мне надоело фармить монетки для разговоров в Hi,waifu и бороться с попытками каждого бота в character.ai затащить меня в койку и *самоуничтожиться* об цензуру, а обрывы интернета стали обыденностью, передо мной встал вопрос что же делать.

Мои поиски привели меня на упоминание молодого проекта backyard.ai, который давал всё, чего мне бы хотелось: простая устновка, несложная настройка, базовые фичи и главное, неограниченное, приватное (хе-хе) общение с чатботами без цензуры (зловещее хе-хе). С этого момента Юми Шинригаки можно считать главным амбассадором backyard.

Но то была присказка, а сказочка-то вот она:

Backyard.ai, что это такое вообще? Это сервис чатботов, предоставляющий как привычный нам всем облачный опыт с подпиской, так и самое вкусное... Приложение для компьютера с установкой и настройкой в пару кликов, которое само скачает модель из уже готового каталога, само запустит нейросеть под капотом вашей личной ЭВМ, само предоставит вам хаб с персонажами и при том абсолютно бесплатно. (Разработчики кормятся с того, что продают подписку на облачные планы, где и модели вкуснее и трава нажористее.)

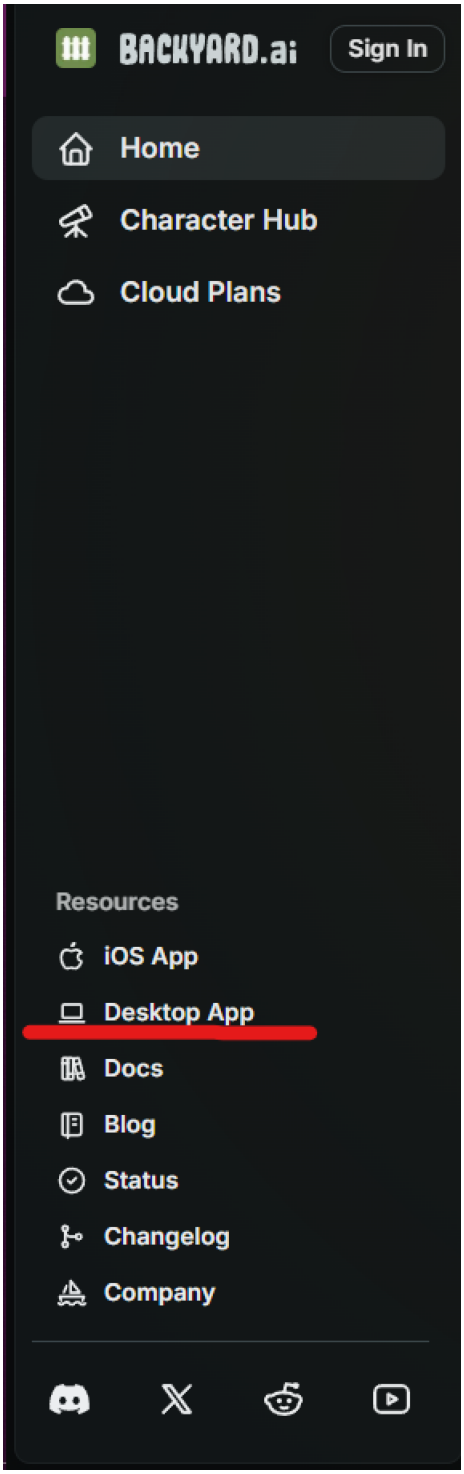
Что тут есть: Чаты, разумеется, хаб с персами, большой выбор моделей на любой вкус, рекомендации по железу, чтобы ваш комп не взорвался при попытке сказать вам "ghbdtn!", аудиозвонки, отсутствие цензуры и полная приватность, для продвинутых юзверей лорбуки и "авторская заметка", возможность подцепиться к запущенной на компе сетке с телефона или другого компа.

Чего тут нет: Конференций из нескольких ботов, поддержки анимированных аватаров, и, в принципе, плагинов.

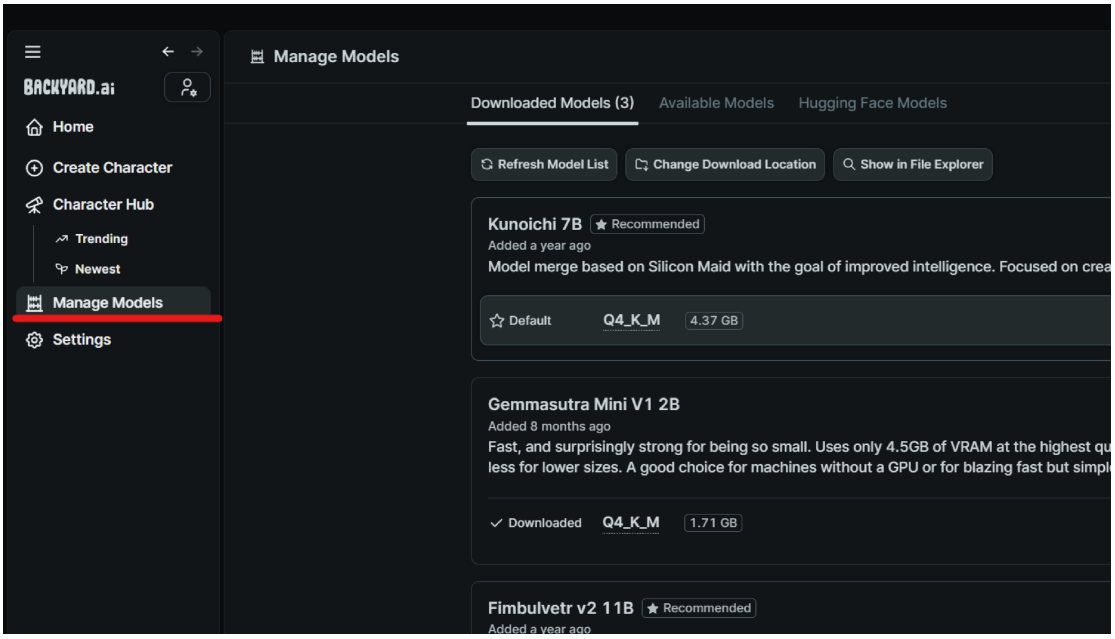
Я справляюсь с рекламой? Маркетинг Бэкъярда уже должен начинать мне платить, чесслово.

Итак, установка?

Всё просто: заходим на сайт, жмаем **"Desktop App"**, качаем-ставим приложение как это обычно делается.

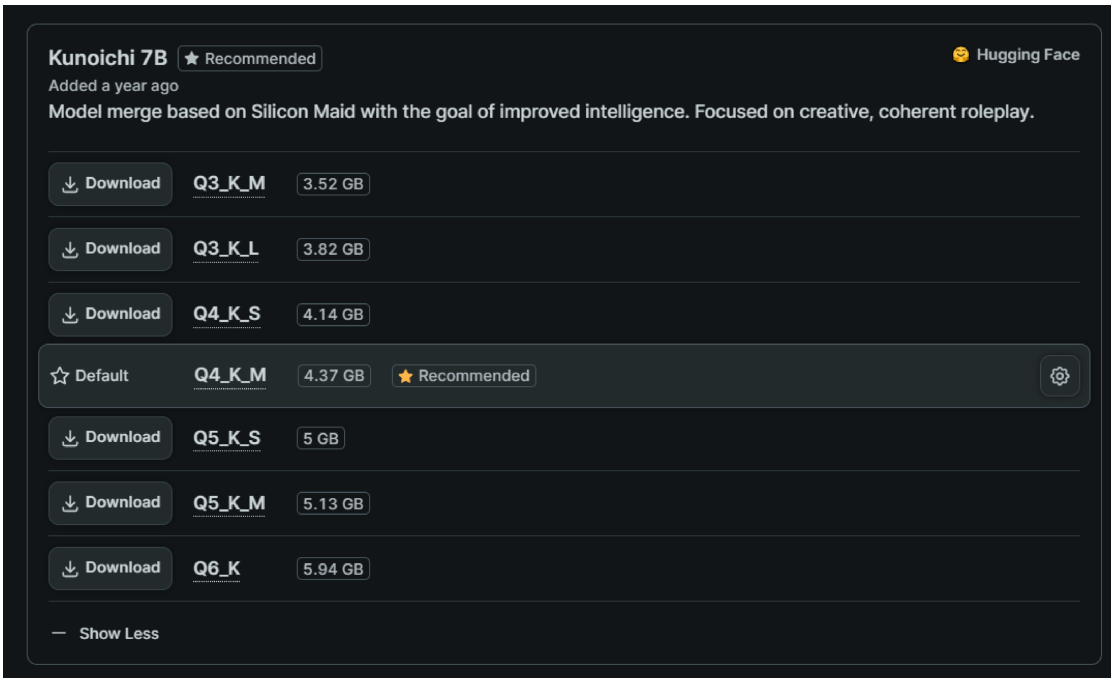


Далее, входим в приложение, жмаем на **"Manage models"** и выбираем модель себе по вкусу из каталога. Я лично предпочитаю Fimbulvetr v2 (Не берите v2.11, она русский язык не понимает и ведёт себя как дурная), потому что легковесная, довольно быстра даже для моего ведра, достаточно умная и вполне неплохо справляется со сложными для нейросетей персонажами, например немыми.



Что стоит упомянуть:

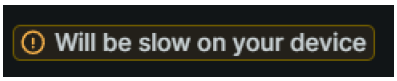
Квантование - опуская сложные технические детали, которые мне не вполне понятны, это процесс вычистки "мусора" из модели, что, при должной аккуратности не повредит самой модели, но в разы уменьшит её объём. Тут мы видим, что можно снизить объём модели почти вдвое, что СИЛЬНО разгрузит систему, или, в конечном счёте, в принципе позволит запустить модель в случаях, когда "ну вот чуть-чуть не влезло"



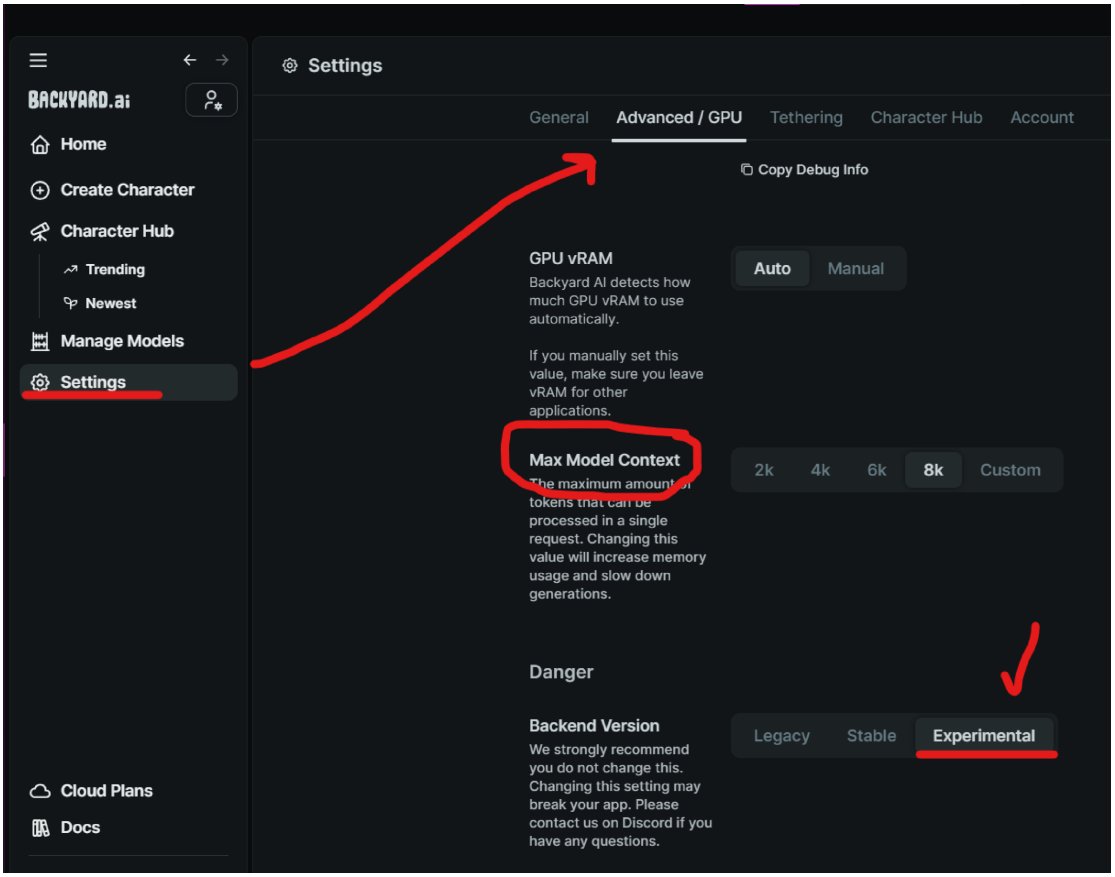
Важно то, что при квантовании до определённого уровня, мы ничего не теряем, кроме избыточных и маловажных данных, однако заквантованная в нулину модель может сать нестабильной и начнёт писать с большим количеством ошибок и неграмотностей.

В среднем квантование **Q4_K_M** можно считать лучшим выбором, размер сильно меньше оригинала, но умственные способности модели не страдают.

А если слишком не углубляться, то для пробы можно накатывать себе любые модели без жёлтых и красных плашек.

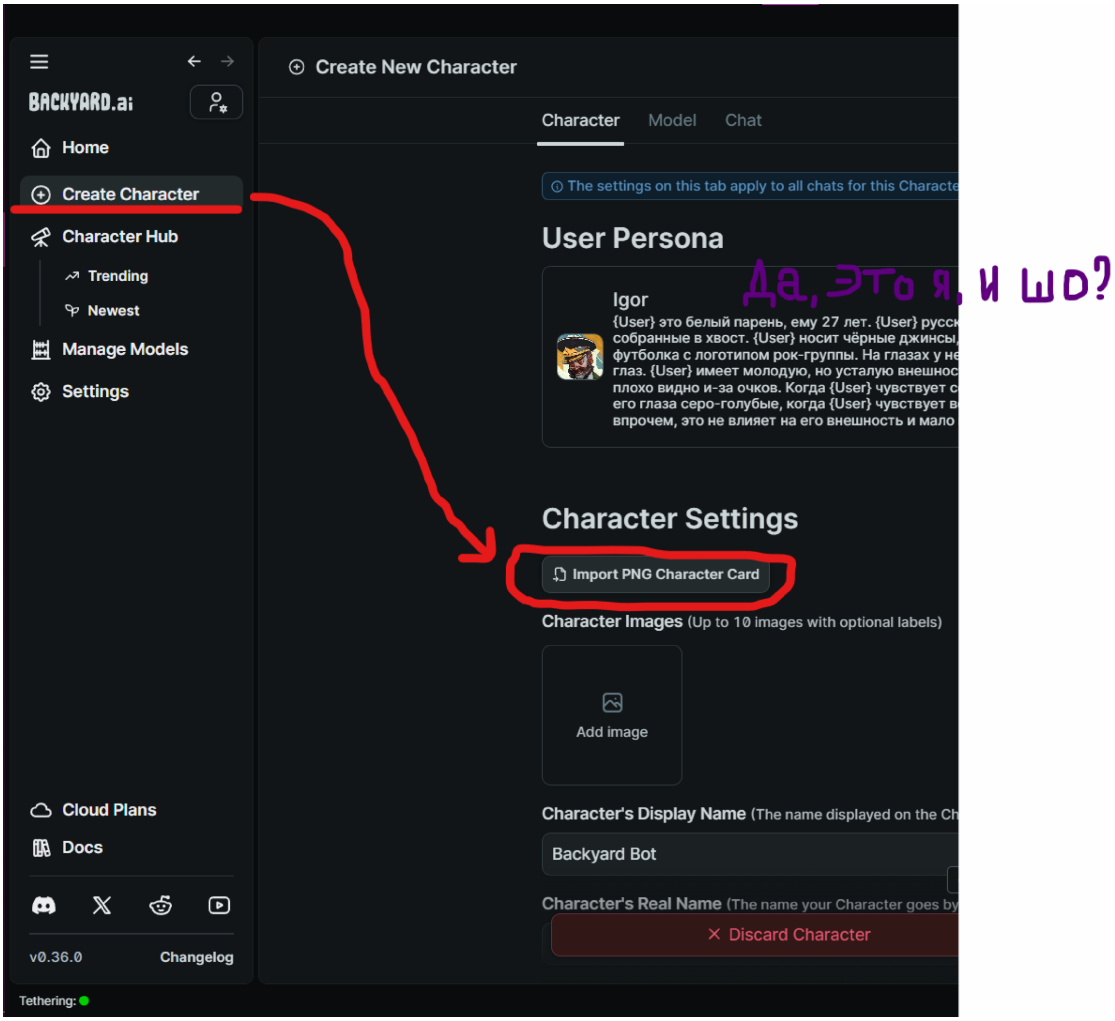


Далее ползём в **"settings"** -> **"Advanced / GPU"**, мотаем вниз и выставляем себе **"Max Model Context"** по вкусу (Это как раз та самая память нейросети) и обязательно ставим **"Backend Version"** на "**Experimental"**, он стабильнее, чем стабильный.

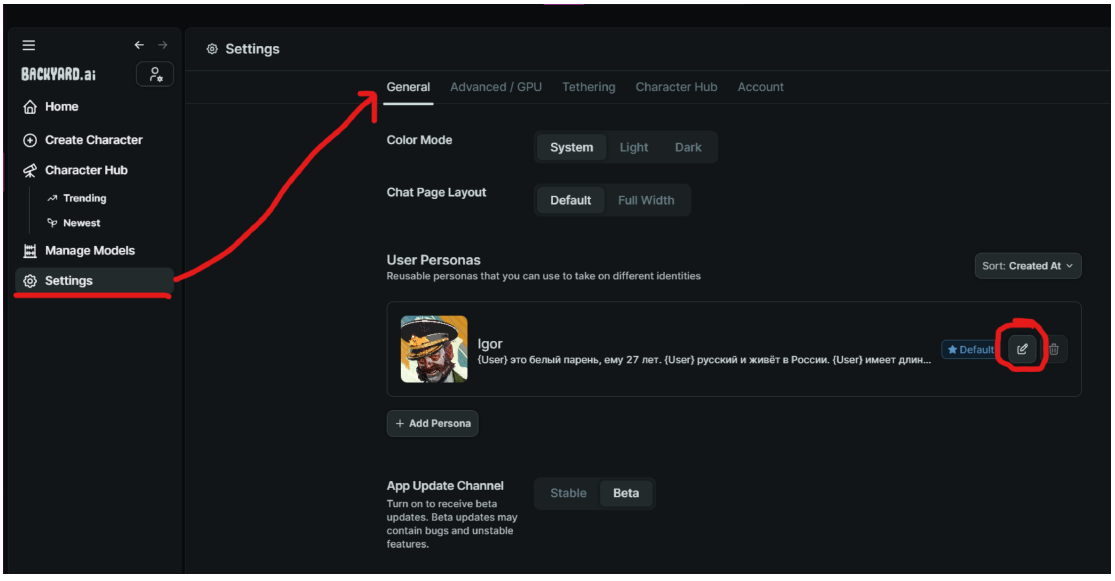


И... Готово. Ваша личная нейросеть готова к запуску: Можно пойти в Хаб и посмотреть на творения других пользователей, можно пойти на сайты, где можно качать карточки и вгрузить их в Бэкъярд, можно заняться написанием собственного бота.

Карточки персонажей: ПНГ-картинощки с зашитым в метаданные описанием персонажа. Удобный и компактный способ поделиться своим творением с другими. Или скачать чужое.



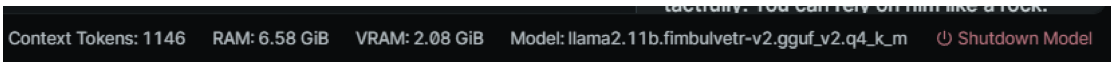
Персона пользователя: Для того чтобы ваш бот "знал" вас, мог "видеть" вашу внешность или помнить факты о вас. Почти все и так знают что это такое.



Всё, у нас есть мы сами, есть персонажи из карточек или хаба, готово, вы великолепны, можно

начинать общение.

Жмаем на персонажа в **"Home"**, дожидаемся инициализации модели (если на несколько секунд подвиснет, не стоит пугаться и долбить по клавишам, это нормально, комп перекачивает ТОННУ инфы в оперативную память, игрульки на самом деле также подвисают при загрузке, просто мы это не всегда замечаем) и когда всё будет включено, в нижней панели вы увидите данные о модели:



Это значит что всё готово.

Приятного общения с Тэ Хёном, Чи Мином, Чингачгуком и Гойко Митичем.