Планы по дообучению GigaChat:

25.04.25 (сегодня) получен доступ в Разделяемое хранилище Hadoop. Разделяемое хранилище Hadoop требуется для сохранения LoRA-адаптера, который будет получен после дообучения Gigachat.

Сегодня (25.04.25) будет запрошен доступ к Datalab Pro – версии Datalab с увеличенным объемом памяти, которое должно позволить сохранить и дообучать GigaChat 2 Max.

После получения доступа к Datalab Pro планируется выполнить дообучение GigaChat 2 Max – если технических проблем (например, автоматическое погашение процесса вследствие неоптимальной загрузки коммунальных ресурсов или каких-либо еще) не возникнет, то дообучение GigaChat 2 Max на собранном датасете, по предварительной оценке, может быть выполнено не более, чем за 2 рабочих дня.

Также следует отметить, что после 16 мая вследствие перевода в другое подразделение, возможно, потребуется повторное получение доступов к Разделяемому хранилищу Hadoop, Datalab Pro, Bitbucket и пр.

Полученный в результате дообучения GigaChat 2 Max LoRA-адаптер будет сохранен на Разделяемое хранилище Hadoop.

Верификация успешности дообучения GigaChat 2 Max может быть выполнена с помощью редакторов на компьютере, на котором выполнялось дообучение (компьютер Акимова Д.А.), только в случае определения, каким образом можно выполнять инференс GigaChat 2 Max аналогично тому, как выполняется инференс при запросе по API или на сайте/телеграм-боте GigaChat – на текущий момент инференс со всех испытанных в Datalab моделей GigaChat значительно отличается в худшую сторону по сравнению с привычным инференсом Gigachat по API или на сайте/телеграм-боте GigaChat (и на текущий момент не удалось найти контакты, которые бы смогли дать ответ, почему так, и как привести инференс GigaChat в Datalab к привычному). Таким образом, если этот вопрос не будет решён, то единственным решением для верификации успешности дообучения GigaChat 2 Max представляется вывод LoRA-адаптера в ПРОМ в SBC в качестве дополнительного (не основного) тестируемого варианта, на который можно переключаться в настройках.

При этом упомянутый в абзаце выше вывод LoRA-адаптера в ПРОМ в SBC может быть выполнен только после того, как в банке будет реализован механизм инференса GigaChat по стандартному тарифу при добавлении к нему LoRA-адаптера (т.е. без оплаты мощностей для развертывания отдельной модели GigaChat с LoRA-адаптером, как сейчас). Данный механизм, по словам Гайнулина И., должен быть реализован во Q2 2025 (в последнем разговоре с ним он не давал более точных оценок, когда это может быть сделано).

В случае недостаточного качества ответов модели после дообучения представляется целесообразным повторно выполнить дообучение при иных гиперпараметрах (в первую очередь, увеличить ранг LoRA и/или количество эпох), а также, возможно, выполнить Preference fine-tuning. Возможность проведения экспериментов и их количество (подбор оптимальных гиперпараметров может быть ресурсозатратным) определяется стоимостью дообучения GigaChat, что может потребовать уточнения (на текущий момент получен довольно размытый ответ от Гудкова П., что стоимость дообучения «размазывается» на кластер без указания более точных цифр).

Вышеуказанные планы могут быть представлены на следующей схеме:

Планы по развитию сервиса парсера новостей (также указаны в презентации по парсеру новостей, направленной 16.04.25):

* доработка и оптимизация сервиса по замечаниям от редакторов, формирующих бизнес-дайджесты (формирование следующего бизнес-дайджеста в июне 2025);
* переход (в случае необходимости) на open-source или отечественные решения;
* реализация мультиагентной системы (в случае запроса от редакторов на автоматическое формирование множества запросов с последующей обработкой результатов запросов и составления итогового текста);
* автоматизация (полная или частичная) создания макета бизнес-дайджеста.