Вихляев А.С., Моисеев Ю.Б. Выбор и производительность локальных больших языковых моделей в задачах автоматической интерпретации психометрических данных// Международный журнал информационных технологий и энергоэффективности. – 2026. – Т. 11 № 6(68) Ч.1 с. 4–10

Международный журнал информационных технологий и энергоэффективности

Сайт журнала:
http://www.openaccessscience.ru/index.php/ijcse/



УДК 004.89:159.9.072

ВЫБОР И ПРОИЗВОДИТЕЛЬНОСТЬ ЛОКАЛЬНЫХ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ В ЗАДАЧАХ АВТОМАТИЧЕСКОЙ ИНТЕРПРЕТАЦИИ ПСИХОМЕТРИЧЕСКИХ ДАННЫХ

[1]Вихляев А.С., [2]Моисеев Ю.Б.
ФГАОУ ВО "МОСКОВСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ" , Москва, Россия (107023, город Москва, Большая Семёновская ул., д. 38), e-mail: [1]privetkomandir@yandex.ru, [2]ybmn@rambler.ru

В работе систематизируются критерии выбора и производительностные характеристики локальных больших языковых моделей применительно к задаче автоматической интерпретации структурированных результатов психометрического тестирования. Задача требует одновременного выполнения нескольких ограничений: локальной обработки данных согласно требованиям российского законодательства о персональных данных, работы на типовом оборудовании рабочей станции, достаточного качества выходного текста. Рассмотрены классы открытых моделей семейства LLaMA, методы посттренировочной квантизации (GPTQ, AWQ, k-quants формата GGUF) и runtime-окружения локального инференса (llama.cpp, Transformers, vLLM). Приведены иллюстративные оценки производительности моделей размером 7–13 миллиардов параметров на CPU и потребительских GPU. Сформулированы рекомендации по выбору конфигурации в зависимости от доступного оборудования и требований к времени отклика.

Ключевые слова: Большие языковые модели; локальный инференс; квантизация; LLaMA; GGUF; защита персональных данных; психометрическая интерпретация; профессиональный отбор.

SELECTION AND PERFORMANCE OF LOCAL LARGE LANGUAGE MODELS IN AUTOMATED INTERPRETATION OF PSYCHOMETRIC DATA

[1]Vikhlyaev A.S., [2]Moiseev Yu.B.
MOSCOW POLYTECHNIC UNIVERSITY, Moscow, Russia (107023, Moscow, Bolshaya Semenovskaya ul., 38,), e-mail: [1]privetkomandir@yandex.ru, [2]ybmn@rambler.ru

The paper systematises the criteria for choosing local large language models and their performance characteristics in the task of automated interpretation of structured psychometric test results. The task imposes several constraints: local data processing as required by the Russian regulations on personal data, operation on conventional workstation hardware, and sufficient output text quality. The work reviews open model families based on LLaMA, post-training quantisation methods (GPTQ, AWQ, k-quants of the GGUF format), and local inference runtimes (llama.cpp, Transformers, vLLM). Indicative performance estimates are given for 7–13 billion parameter models on CPU and consumer-grade GPUs. Configuration recommendations are formulated as a function of the available hardware and the response-time requirements.

Keywords: Large language models; local inference; quantisation; LLaMA; GGUF; personal data protection; psychometric interpretation; professional selection.

Введение

Цифровизация процедур психологического отбора привела к росту объёма структурированных данных, поступающих от каждой методики тестирования: помимо итоговых баллов фиксируются временные метрики, профили выполнения, паттерны

переключения между заданиями. Объём этих данных делает ручную интерпретацию специалистом трудоёмкой, особенно в массовых процедурах врачебно-лётной экспертизы и кадрового отбора в операторские профессии.

Большие языковые модели на основе трансформерной архитектуры [8] позволяют обрабатывать структурированные числовые данные и порождать развёрнутый текст интерпретации. Прямое применение таких моделей через облачные API в задачах ведомственного отбора затруднено нормативно: Федеральный закон № 152-ФЗ предписывает локализацию обработки персональных данных [3]. Психодиагностические данные кандидатов на лётные и операторские должности подпадают под усиленные требования к защите, и приемлемой архитектурой служит локальная обработка, при которой и данные, и модель находятся в контуре заказчика.

Развитие открытых моделей семейства LLaMA [7] и методов посттренировочной квантизации [2, 5] сделало локальный инференс на типовом оборудовании технически возможным. Возникает инженерная задача выбора конфигурации модели и runtime-окружения, обеспечивающих приемлемое качество интерпретации и время отклика, совместимое с работой специалиста. Компромиссы между размером модели, уровнем квантизации, типом оборудования и качеством вывода имеют многомерный характер.

На основе анализа литературы поставлена цель работы: систематизировать критерии выбора локальной большой языковой модели для задачи автоматической интерпретации психометрических данных и оценить производительностные характеристики типовых конфигураций. Исходя из цели исследования решались следующие задачи: 1) проанализировать применимые классы открытых языковых моделей и оценить их пригодность для генерации развёрнутой психометрической интерпретации в условиях ведомственного применения; 2) сопоставить методы посттренировочной квантизации и runtime-окружения локального инференса в части соотношения качества и ресурсных требований; 3) сформулировать функциональные требования к программной системе автоматической интерпретации, согласованные с нормами защиты персональных данных и с практикой профотбора в операторские профессии; 4) привести иллюстративные оценки производительности типовых конфигураций «модель — квантизация — оборудование» и сформулировать рекомендации по их выбору.

1. Литературный обзор

Современные большие языковые модели опираются на трансформерную архитектуру, предложенную Vaswani с соавт. [8]. Механизм самовнимания позволяет обрабатывать последовательность токенов параллельно и улавливать зависимости произвольной длины, что сделало возможным масштабирование моделей до десятков и сотен миллиардов параметров. Открытые модели меньшего размера заполнили нишу локального применения: Touvron с соавт. описали семейство LLaMA с моделями от 7 до 65 миллиардов параметров и показали, что модель на 13 миллиардов параметров превосходит модели семейства GPT-3 размером 175 миллиардов на большинстве сопоставительных задач [7]. Для прикладных задач необязательно использовать самую крупную модель; компактная модель с грамотно подобранным обучающим корпусом обеспечивает сопоставимое качество при кратно меньших ресурсных требованиях.

Параллельно сформировался класс методов посттренировочной квантизации. Метод LLM.int8(), предложенный Dettmers с соавт., сократил объём памяти вдвое без значимой деградации качества за счёт сочетания векторно-блочной квантизации основных весов и сохранения в полной точности доли «выбросов» [2]. Frantar с соавт. показали, что метод GPTQ обеспечивает квантизацию до 3–4 бит на вес с минимальным ростом перплексии [6]. Lin с соавт. в работе о методе AWQ установили, что защиты заслуживает лишь около одного процента «значимых» весов, и для остальных допустимо понижение точности до четырёх бит [5]. Практическая применимость на потребительском оборудовании была обеспечена развитием runtime-окружений; проект llama.cpp ввёл формат GGUF и семейство k-квантов, оптимизированных для CPU и архитектуры Apple Silicon [9].

Психодиагностический контекст применения опирается на сложившуюся практику профессионального отбора в операторские профессии [1] и общие требования к интерпретации психометрических данных [4]. Цифровая среда тестирования сама по себе вносит дополнительную вариативность в исходные данные [10], что усиливает требования к качеству автоматической интерпретации. Отдельная задача возникает в специализированных сценариях профотбора лётного состава, в частности при оценке функционального состояния после пережитых экстремальных событий [11].

2. Выбор модели, квантизации и runtime-окружения

К рассматриваемой системе предъявляются четыре класса требований: локальность обработки персональных данных в соответствии с № 152-ФЗ [3]; совместимость с типовым оборудованием рабочей станции специалиста (CPU 8–16 ядер, оперативная память 16–32 ГБ, опционально потребительский графический процессор с 8–16 ГБ видеопамяти); качество выходного текста, достаточное для использования в составе экспертной процедуры; время отклика порядка 5–30 секунд на одного испытуемого. Перечисленные требования сводятся к инженерной задаче подбора модели, метода квантизации и runtime, при которых одновременно соблюдены границы по памяти, по времени отклика и по качеству вывода.

Размер модели образует первый параметр выбора. Модели на 1–3 миллиарда параметров обеспечивают высокую скорость инференса даже на CPU (порядка 30–50 токенов в секунду), но дают ограниченное качество на задачах структурированного следования инструкции. Модели на 7–8 миллиардов параметров характеризуются в инженерной практике как оптимум для локальных задач широкого профиля: они удерживают структуру длинного промпта, корректно обращаются с числовыми входными данными при условии явной разметки, выполняют инструкции по формату ответа; объём памяти для квантизованной версии лежит в диапазоне 4–8 ГБ. Модели на 13 миллиардов параметров дают заметный прирост качества на задачах многошагового рассуждения и при наличии потребительского графического процессора с 12–16 ГБ видеопамяти работают с приемлемой скоростью. Рассматриваются модели с инструкционным дообучением (instruction-tuned), а не базовые foundation-модели: последние обучены на задаче продолжения текста и на инструкции реагируют непредсказуемо.

Квантизация образует второй параметр выбора. Веса трансформерной модели не равноценны: малая часть весов доминирует в выходе, и именно их точность определяет качество модели в целом [2, 5]. На уровне 5 бит на вес (Q5_K_M в формате GGUF) деградация

качества относительно полной точности FP16 не превышает 1 % по перплексии; уровень 4 бита на вес (Q4_K_M) даёт сокращение памяти приблизительно в четыре раза при умеренной деградации и остаётся пригодным для большинства прикладных применений. Уровни ниже Q4 применимы только при жёстких ресурсных ограничениях и сопровождаются заметным ростом ошибок формата вывода. Выбор runtime-окружения замыкает связку: для рабочих станций без выделенной видеокарты и для Apple Silicon естественным выбором служит llama.cpp с форматом GGUF [9]; для потребительских NVIDIA-видеокарт 8–16 ГБ применимы Transformers с пакетами AutoGPTQ или AutoAWQ либо специализированный runtime ExLlamaV2; для серверного развёртывания применим сервер vLLM с механизмом PagedAttention.

3. Производительностные характеристики

Производительность локального инференса характеризуется тремя параметрами: скоростью генерации (токенов в секунду на этапе декодирования), временем до первого токена (обработка входного промпта) и объёмом потребляемой памяти. Сводные иллюстративные показатели для типовых конфигураций приведены в Таблице 1; числа отражают типичные значения на современном потребительском оборудовании и зависят от длины контекста, сборки runtime, версии драйверов, теплового режима оборудования.

Таблица 1 - Иллюстративная производительность локального инференса для моделей семейства LLaMA и совместимых архитектур

Размер модели	Квантизация	Оборудование	Память	Скорость, ток./с
3 млрд	Q4_K_M	CPU 8 ядер	≈ 2 ГБ	30–50
7–8 млрд	Q4_K_M	CPU 8 ядер	≈ 5 ГБ	8–12
7–8 млрд	Q4 (GPTQ/AWQ)	GPU 8–12 ГБ	≈ 5 ГБ	40–60
13 млрд	Q4_K_M	CPU 16 ядер	≈ 8 ГБ	4–7
13 млрд	Q4 (GPTQ/AWQ)	GPU 12–16 ГБ	≈ 9 ГБ	25–40
7–8 млрд	Q4_K_M	Apple M-серия	≈ 5 ГБ	25–40

Из приведённых данных следуют практические наблюдения. На CPU модели размером 7–8 миллиардов параметров с квантизацией Q4_K_M обеспечивают скорость генерации 8–12 токенов в секунду, что при типичной длине ответа 300–500 токенов даёт интерпретацию одного протокола за 30–60 секунд и укладывается в требование по времени отклика. Переход на потребительский графический процессор увеличивает скорость в 4–6 раз. Архитектура Apple Silicon с унифицированной памятью обеспечивает производительность, сопоставимую с дискретным графическим процессором, при существенно меньшем энергопотреблении.

4. Применение в задачах профотбора и ограничения

Прикладная задача автоматической интерпретации психометрических данных в профотборе операторов предполагает поддержку специалиста, а не его замену. Интерпретация, сгенерированная локальной языковой моделью, предъявляется специалисту в качестве заготовки и подлежит экспертной проверке перед включением в итоговое заключение. Соотнесение с требованиями российского законодательства о персональных данных при локальной архитектуре оказывается прямым: и данные, и средство обработки находятся в контуре заказчика, передачи внешним сервисам не происходит, состав обрабатываемых данных минимизируется до необходимого для интерпретации [3].

Один из практически значимых сценариев применения связан с задачами психологического сопровождения лётного состава, в особенности с регулярной переоценкой функционального состояния после пережитых экстремальных событий. В работе Ю.Б. Моисеева о психологических особенностях лётчиков, связанных с катапультированием, фиксируется, что после такого события у лётчика регулярно отмечаются изменения характеристик внимания, оперативной памяти и быстроты принятия решения [11]. В этой задаче автоматизация интерпретации даёт ощутимый выигрыш: специалист получает заготовку отчёта, требующую только верификации и экспертной коррекции.

Ограничения подхода: локальные модели среднего размера допускают фактические галлюцинации и пропуски значимых шкал, что требует обязательной экспертной проверки; воспроизводимость требует фиксации точных параметров генерации в протоколе обработки; обновление моделей и runtime-окружений предъявляет требования к версионированию и валидации новых сборок перед допуском в эксплуатацию.

Заключение

В работе систематизированы критерии выбора локальной большой языковой модели для задачи автоматической интерпретации психометрических данных и приведены иллюстративные оценки производительности типовых конфигураций. Показано, что задача допускает практическое решение на современном потребительском оборудовании при использовании инструкционно дообученных моделей размером 7–13 миллиардов параметров, посттренировочной квантизации до 4–5 бит на вес и выборе runtime-окружения сообразно доступному оборудованию: llama.cpp для CPU и Apple Silicon, Transformers или ExLlamaV2 для потребительских NVIDIA-видеокарт, vLLM для серверного развёртывания. Локальная архитектура удовлетворяет нормативным требованиям по защите персональных данных, изложенным в Федеральном законе № 152-ФЗ.

Работа имеет прикладное значение для проектирования программных систем психологического сопровождения профессиональной деятельности, в которых требуется сочетание автоматизации интерпретации, локальности обработки данных и работы в условиях ограниченных вычислительных ресурсов. Прикладной фокус задаётся практикой профессионального отбора в операторские профессии, в том числе в авиационной медицине и психологии. Сценарий регулярной переоценки функционального состояния лётного состава после пережитых экстремальных событий [11] относится к задачам, в которых соотношение качества и трудоёмкости интерпретации особенно чувствительно. Дальнейшая работа предполагает эмпирическую оценку качества интерпретаций, генерируемых конкретными моделями на размеченных корпусах.

Список литературы

1. Бодров В.А., Лукьянова Н.Ф. Психологический отбор лётчиков и космонавтов. — М.: Наука, 1984. — 264 с.
2. Dettmers T., Lewis M., Belkada Y., Zettlemoyer L. LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale // Advances in Neural Information Processing Systems 35. — 2022. — pp. 30318–30332. — DOI: 10.48550/arXiv.2208.07339.
3. О персональных данных: Федеральный закон от 27.07.2006 № 152-ФЗ (с изм.) // Собрание законодательства Российской Федерации. — 2006. — № 31 (ч. 1). — Ст. 3451.
4. Бурлачук Л.Ф. Психодиагностика: Учебник для вузов. 2-е изд. — СПб.: Питер, 2017. — 384 с.
5. Lin J., Tang J., Tang H., Yang S., Dang X., Gan C., Han S. AWQ: Activation-aware Weight Quantization for LLM Compression and Acceleration // Proceedings of Machine Learning and Systems 6 (MLSys 2024). — 2024. — DOI: 10.48550/arXiv.2306.00978.
6. Frantar E., Ashkboos S., Hoefler T., Alistarh D. GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers // Proceedings of the 11th International Conference on Learning Representations (ICLR 2023). — 2023. — DOI: 10.48550/arXiv.2210.17323.
7. Touvron H., Lavril T., Izacard G. et al. LLaMA: Open and Efficient Foundation Language Models. — 2023. — DOI: 10.48550/arXiv.2302.13971.
8. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I. Attention Is All You Need // Advances in Neural Information Processing Systems 30. — 2017. — pp. 6000–6010. — DOI: 10.48550/arXiv.1706.03762.
9. Gerganov G. llama.cpp: LLM Inference in C/C++ [Электронный ресурс]. — URL: https://github.com/ggml-org/llama.cpp.
10. Mead A.D., Drasgow F. Equivalence of computerized and paper-and-pencil cognitive ability tests: A meta-analysis // Psychological Bulletin. — 1993. — Vol. 114, No 3. — pp. 449–458. — DOI: 10.1037/0033-2909.114.3.449.
11. Моисеев Ю.Б. Психологические особенности лётчиков, связанные с катапультированием // Военная авиационная психология: Сборник лекций / под ред. В.А. Пономаренко. — М.: Перо, 2021. — С. 241–261.

References

1. Bodrov V.A., Lukyanova N.F. Psychological Selection of Pilots and Cosmonauts. Moscow: Nauka, 1984, p.264
2. Dettmers T., Lewis M., Belkada Y., Zettlemoyer L. LLM.int8(): 8-bit Matrix Multiplication for Transformers at Scale // Advances in Neural Information Processing Systems 35. — 2022. — pp. 30318–30332. — DOI: 10.48550/arXiv.2208.07339.
3. On Personal Data: Federal Law of July 27, 2006, No. 152-FZ (as amended) // Collected Legislation of the Russian Federation. — 2006. — No. 31 (Part 1). — Art. 3451.
4. Burlachuk L.F. Psychodiagnostics: Textbook for Universities. 2nd ed. — St. Petersburg: Piter, 2017. — p.384
5. Lin J., Tang J., Tang H., Yang S., Dang X., Gan C., Han S. AWQ: Activation-aware Weight Quantization for LLM Compression and Acceleration // Proceedings of Machine Learning and Systems 6 (MLSys 2024). — 2024. — DOI: 10.48550/arXiv.2306.00978.

6. Frantar E., Ashkboos S., Hoefler T., Alistarh D. GPTQ: Accurate Post-Training Quantization for Generative Pre-trained Transformers // Proceedings of the 11th International Conference on Learning Representations (ICLR 2023). - 2023. - DOI: 10.48550/arXiv.2210.17323.
7. Touvron H., Lavril T., Izacard G. et al. LLaMA: Open and Efficient Foundation Language Models. - 2023. - DOI: 10.48550/arXiv.2302.13971.
8. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I. Attention Is All You Need // Advances in Neural Information Processing Systems 30. - 2017. - pp. 6000–6010. — DOI: 10.48550/arXiv.1706.03762.
9. Gerganov G. llama.cpp: LLM Inference in C/C++ [Electronic resource]. — URL: https://github.com/ggml-org/llama.cpp.
10. Mead A.D., Drasgow F. Equivalence of computerized and paper-and-pencil ability cognitive tests: A meta-analysis // Psychological Bulletin. - 1993. - Vol. 114, No. 3. — pp. 449–458. — DOI: 10.1037/0033-2909.114.3.449.
11. Moiseev Yu.B. Psychological characteristics of pilots associated with ejection // Military aviation psychology: Collection of lectures / edited by V.A. Ponomarenko. — Moscow: Pero, 2021. — pp. 241–261.