



Повышение эффективности квантования больших языковых моделей с архитектурой смеси экспертов

Д.Р. Пойманов, ORCID: 0009-0001-5390-915X <poimanovdr@my.msu.ru>

Московский государственный университет имени М.В. Ломоносова,
Россия, 119991, Москва, Ленинские горы, д. 1.

Аннотация. Модели с архитектурой смеси экспертов (Mixture-of-experts, MoE) требуют значительных ресурсов для хранения и выполнения, поэтому их сжатие является актуальной задачей. Посттренировочное квантование весов – один из наиболее эффективных способов уменьшения размера модели, однако существующие методы, например метод GPTQ, при прямом применении к модели MoE работают хуже из-за неравномерной активации экспертов и особенностей маршрутизации. Для устранения указанных недостатков в работе предложены три модификации. Первая – формирование калибровочного набора с повышенной энтропией активаций экспертов за счёт семплирования данных из разных доменов. Вторая – взвешивание гессаана коэффициентами маршрутизатора для более точной оценки вклада экспертов. Третья – введение метрики чувствительности эксперта, позволяющей назначать двухбитное квантование наименее значимым экспертам и трёхбитное – остальным. Эксперименты на модели Qwen3-30B-A3B и Mixtral-8x7b показали, что предложенный метод превосходит существующие аналоги в режиме INT3, а использование метрики чувствительности позволяет понизить среднюю битность для модели до 2.8 с минимальным падением качества генерации текстов. Полученные результаты подтверждают эффективность предложенных модификаций и открывают путь к более компактному развёртыванию MoE-моделей.

Ключевые слова: большие языковые модели; квантование; оптимизация вычислений языковых моделей; сжатие нейросетей; посттренировочное квантование.

Для цитирования: Пойманов Д.Р. Повышение эффективности квантования больших языковых моделей с архитектурой смеси экспертов. Труды ИСП РАН, том 38, вып. 6, часть 1, 2026 г., стр. 123–138. DOI: 10.15514/ISPRAS-2026-38(6)-8.

Enhancing Quantization Efficiency in Mixture-of-Experts Large Language Models

D.R. Poimanov, ORCID: 0009-0001-5390-915X <poimanovdr@my.msu.ru>

Lomonosov Moscow State University,
GSP-1, Leninskie Gory, Moscow, 119991, Russia.

Abstract. Models based on the Mixture-of-Experts (MoE) architecture demand substantial storage and computational resources, making their compression a pressing challenge. Post-training weight quantization is among the most effective means of reducing model size; however, existing methods such as GPTQ exhibit degraded performance when directly applied to MoE models due to imbalanced expert activation and the specific nature of sparse routing. To address these limitations, this work introduces three modifications. First, a calibration dataset with elevated expert activation entropy is constructed by sampling data from diverse domains. Second, the Hessian matrix is weighted by router coefficients to better capture the contribution of individual experts. Third, an expert sensitivity metric is proposed, enabling 2-bit quantization for the least important experts while retaining 3-bit precision for the remainder. Experiments on the Qwen3-30B-A3B and Mixtral-8x7b model demonstrate that the proposed method outperforms existing approaches in the INT3 regime, and the sensitivity metric further reduces the average bit-width to 2.8 with minimal degradation in text generation quality. These findings confirm the effectiveness of the proposed modifications and pave the way toward more compact deployment of MoE models.

Keywords: large language models; quantization; LLM inference optimization; neural networks compression; PTQ methods.

For citation: Poimanov D.R. Enhancing Quantization Efficiency in Mixture-of-Experts Large Language Models. *Trudy ISP RAN/Proc. ISP RAS*, vol. 38, issue 6, part 1, 2026, pp. 123-138 (in Russian). DOI: 10.15514/ISPRAS-2026-38(6)-8.

1. Введение

Большие языковые модели (Large Language Models, LLM) добились впечатляющих успехов в широком спектре задач обработки естественного языка, включая понимание текста, логические рассуждения и генерацию контента, демонстрируя превосходную производительность и адаптивность [1-2]. Однако стремительный рост размеров моделей – современные флагманские LLM насчитывают десятки и сотни миллиардов параметров – оборачивается серьёзными проблемами, связанными с нагрузкой на вычислительные ресурсы и высоким потреблением памяти [3-4]. Эти ограничения становятся критическими при развёртывании моделей на устройствах с ограниченными ресурсами и в сценариях, требующих обработки больших потоков запросов.

В качестве перспективного решения для снижения вычислительных издержек при обучении и инференсе была предложена архитектура «смесь экспертов» (Mixture of Experts, MoE) [5-6]. Подход MoE позволяет наращивать общее количество параметров модели, сохраняя число активных (задействованных для обработки одного токена) параметров на относительно постоянном уровне. Благодаря использованию множества специализированных подсетей – «экспертов» – и механизма разреженной маршрутизации, MoE обеспечивает эффективные вычисления, открывая путь к созданию более ёмких моделей в условиях ограниченной вычислительной мощности.

Тем не менее, даже при использовании MoE-архитектуры размер моделей в памяти остаётся значительным, а объём занимаемого дискового пространства и требуемой оперативной памяти по-прежнему ограничивает их практическое применение. Одним из наиболее действенных способов уменьшения этих требований является квантование – процедура понижения разрядности представления параметров модели. Однако прямое применение стандартных методов посттренировочного квантования, хорошо зарекомендовавших себя на плотных архитектурах, к моделям со смесью экспертов сопряжено с рядом специфических

трудностей, обусловленных неравномерной активацией экспертов и особенностями маршрутизации. Настоящая работа посвящена разработке и экспериментальному обоснованию модификаций метода GPTQ, направленных на преодоление указанных ограничений и повышение эффективности квантования МоЕ-моделей.

2. Основные понятия

2.1 Архитектура «Mixture of Experts»

Идея «Mixture of Experts» восходит к началу 1990-х годов [7], однако реальный прорыв в применении к большим языковым моделям произошёл позже. В период 2020-21 годов Google представил архитектуры GShard [8] и Switch Transformer [9], которые впервые успешно интегрировали МоЕ-слои в глубокие трансформеры.

В традиционной архитектуре языковой модели, представленной в виде трансформера [10], каждый токен проходит через один и тот же слой однонаправленной нейронной сети (Feed-forward network, FFN). При масштабировании модели это приводит к быстрому росту вычислительных затрат. МоЕ предлагает принципиально иной подход: вместо одного универсального FFN-блока используется целый набор «экспертов» – небольших нейронных сетей (обычно таких же FFN-модулей), – а специальный роутер (или маршрутизатор) для каждого входного токена выбирает лишь нескольких из них. В результате модель сохраняет высокую ёмкость (общее число параметров может быть огромным), но для каждого токена активируется лишь малая доля этих параметров, что позволяет удерживать объём вычислений примерно на одном уровне.

Пусть x – вектор входной последовательности для одного токена. Результат вычисления для обычного FFN-слоя можно записать как $y = f(x)$. Однако в случае использования МоЕ FFN-часть заменяется на набор экспертов $f_1, f_2, \dots, f_N$ и роутер G . Роутером, как правило, является линейный слой, который вычисляет вектор значений $h(x) = W_r x + b_r$, где W_r, b_r – обучаемые параметры. После применяется функция softmax по K наибольшим значениям вектора для вычисления вектора коэффициентов МоЕ слоя:

$$G(x) = \text{softmax}(\text{TopK}(h(x), K))$$

Выход МоЕ-слоя для токена представляет собой взвешенную сумму выходов только тех экспертов, которые были выбраны роутером:

$$y = \sum_{i \in \text{TopK}(h(x), K)} G_i(x) f_i(x)$$

Таким образом, для каждого токена внутри каждого МоЕ-слоя используется свой набор из K экспертов. Например, одной из рассматриваемых в работе моделей является Qwen3-30B-A3B [11], для которой $K = 8, N = 128$, то есть доля активных параметров составляет примерно 6.25%. Именно это и даёт основной выигрыш в вычислительной эффективности.

2.2 Сжатие больших языковых моделей

Как было сказано ранее, современные LLM, в том числе построенные по архитектуре МоЕ, насчитывают миллиарды и десятки миллиардов параметров. Хранение таких моделей в стандартном 16- или 32-битном числовом формате требует десятков и сотен гигабайт дискового пространства. Для снижения требований к памяти и дисковому хранилищу разработан ряд подходов, объединяемых общим термином «сжатие моделей». Среди них можно выделить такие как:

- **дистилляция знаний** [12-13] – обучение компактной модели «ученика» на выходных распределениях большой модели «учителя»;
- **прореживание** (pruning) [14-15] – удаление наименее значимых весов или целых структурных блоков (нейронов, каналов) из графа вычислений;

- **матричная факторизация** [16-17] – представление матриц весов в виде произведения матриц меньшего ранга;
- **квантование** [4,18] – уменьшение разрядности представления параметров и/или активаций.

В данной работе все внимание сконцентрировано на алгоритмах квантования, поскольку среди перечисленных методов на сегодняшний день они являются наиболее эффективным и универсальным инструментом сжатия LLM. Основные преимущества обусловлены тем, что квантование не требует длительного дообучения (в отличие от дистилляции знаний), позволяет сократить размер модели в 2–4 раза без катастрофической потери качества и напрямую ускоряет вычисления на совместимом аппаратном обеспечении. Более того, в промышленных решениях квантование получило особенно широкое распространение благодаря тому, что легко сочетается с другими подходами по уменьшению размера моделей.

2.3 Квантование параметров

2.3.1 Определение и свойства

Более формально под квантованием в контексте нейронных сетей понимается преобразование чисел из высокоточного представления (например, FP32, FP16) в дискретное низкоразрядное представление (обычно INT4, INT3, и т.д.). В данном исследовании рассматривается наиболее распространенное равномерное квантование (INT), которое для скалярного вещественного значения r определяется отображением:

$$q = \text{clamp}(\text{round}(\frac{r-z}{s}), 0, 2^b - 1),$$

где b – число бит; s, z – параметры квантования (могут быть обучаемыми или константными); $\text{round}(\cdot)$ – операция округления до ближайшего целого. Функция $\text{clamp}(\cdot)$ ограничивает входное значение заданным интервалом $[0, 2^b - 1]$: если значение выходит за нижнюю границу, возвращается 0; если за верхнюю, то округляется до $2^b - 1$; в остальных случаях возвращается само значение. Как пример, в случае, когда $b = 4$ полученное значение q имеет формат INT4.

Восстановление приближённого исходного значения (деквантование) выполняется обратным преобразованием:

$$r_{\text{deq}} = s \cdot q + z$$

Эта процедура необходима для того, чтобы привести квантованные INT веса к формату FP, в котором далее производится матричное умножение на активации и выполняются остальные арифметические операции. Следует отметить, что параметры квантования s, z также подлежат хранению в памяти для последующего деквантования, причём каждый из них представляется в формате FP16. Вследствие этого квантование применяется не к отдельным весам, а к группам параметров, что позволяет сократить объём дополнительно хранимой информации и, как следствие, уменьшить общий размер модели в памяти.

2.3.2 Методы посттренировочного квантования

В задачах сжатия нейросетевых моделей выделяют два основных класса методов квантования: **квантование с дообучением** [19-20] (Quantization-Aware Training, QAT) и **посттренировочное квантование** [4,18,21] (Post-Training Quantization, PTQ). При использовании QAT модель проходит дополнительный цикл обучения, в ходе которого симулируется влияние квантования, что позволяет весам адаптироваться к пониженной точности и минимизировать потери качества. Такой подход, как правило, даёт наилучшие результаты с точки зрения итоговой точности, однако требует полного набора обучающих

данных, доступа к полному вычислительному графу модели и значительных вычислительных ресурсов, сопоставимых с затратами на исходное обучение.

В отличие от QAT, посттренировочное квантование применяется к уже обученной модели без дополнительной настройки параметров или с минимальной калибровкой на небольшом подмножестве данных. Такой подход не требует ни полного набора исходных обучающих данных, ни дорогостоящего переобучения, что делает его практически единственно возможным решением для применения к большинству современных LLM в условиях ограниченных вычислительных ресурсов. В настоящей работе фокус сознательно сделан именно на методах посттренировочного квантования.

3. Постановка исследовательской задачи

Переходя от общего понимания квантования как процедуры сжатия весов модели к более узкой и практически значимой постановке, сформулируем исследовательскую задачу данной работы. Соединение введенных ранее понятий – архитектуры MoE и методов посттренировочного квантования – позволяет очертить контуры проблемы, которая на первый взгляд выглядит достаточно прямолинейно. В самой общей формулировке задача может быть сведена к простому требованию: необходимо сжать веса MoE-модели при помощи методов посттренировочного квантования настолько сильно, насколько это возможно, не допустив при этом падения качества генерации текста. Такая постановка, однако, является чрезмерно широкой: для перехода к конкретному техническому решению требуется детальное изучение существующего арсенала методов, представленных в литературе, и всесторонняя оценка того, насколько имеющиеся подходы способны учесть специфику MoE-архитектур.

Анализ существующих подходов к квантованию MoE-моделей, подробный обзор которых будет представлен в следующем разделе, показал, что наиболее зрелые и хорошо изученные подходы разрабатывались и отлаживались преимущественно на обычных архитектурах трансформеров. Их эффективность подтверждена многочисленными экспериментами на моделях семейства LLaMA [22], где активация весов MLP (multilayer perceptron) блока происходит равномерно, а каждый параметр задействуется при обработке любого входного токена. В то же время попытки прямого переноса этих методов на MoE-модели без учёта их специфики демонстрируют ошутимое снижение качества, особенно при переходе к низким разрядностям, таким как 3 или 2 бита на вес. Причина такого расхождения кроется в фундаментальном отличии характера использования параметров: в MoE-слое эксперты активируются разреженно и неравномерно, их вклад в итоговое представление модулируется коэффициентами маршрутизатора, а статистические свойства активаций, используемые для оценки чувствительности весов к квантованию, могут существенно варьироваться от эксперта к эксперту.

Таким образом, исследовательская задача данной работы заключается не в изобретении принципиально нового алгоритма квантования, а в целенаправленной модификации и расширении существующих подходов с возможностью учитывания специфики разреженной маршрутизации и, как следствие, обеспечения более высокого качества сжатия при экстремально низких битностях.

4. Обзор существующих подходов

Одной из первых работ, систематически исследующих посттренировочное квантование MoE-моделей, является QuantMoE-Bench [23]. Авторы проводят всесторонний анализ чувствительности различных компонентов архитектуры к понижению разрядности и показывают, что разные части MoE-слоёв требуют разного числа бит для сохранения качества. На основе этих наблюдений предложены эвристики для распределения битности и

методы оценки важности линейных слоёв, учитывающие выбросы активаций. QuantMoE-Bench определяет общие принципы квантования MoE-моделей, но не даёт готового решения, учитывающего особенности конкретного калибровочного набора.

Следующим шагом стала работа MoEQuant [24], в которой внимание сосредоточено на двух типах дисбаланса, присущих MoE: межэкспертном и внутриэкспертном. Для их устранения предложены два метода. Первый формирует калибровочный набор со сбалансированной активацией экспертов, используя кумулятивные вероятности токенов. Второй вводит в процедуру квантования информацию между экспертами и входными примерами. Эксперименты на нескольких MoE-моделях демонстрируют значительный прирост точности по сравнению со стандартным GPTQ.

Гибридный подход MC-MoE [25] объединяет статическое квантование и динамическое прореживание для достижения экстремального сжатия MoE-моделей. MC-MoE не требует дообучения и использует информацию о значимости как экспертов, так и токенов для максимизации степени сжатия при минимальной потере точности. Авторы демонстрируют, что при среднем размере веса 2,54 бита метод сжимает 76,6% модели, теряя в среднем лишь 3,8% точности. Работа показательна тем, что демонстрирует эффективность комбинирования различных методов сжатия для MoE-архитектур.

Таким образом, анализ литературы выявляет тенденцию эволюции от общих методов оценки и установления базовых закономерностей квантования нейросетей с MoE архитектурой к созданию узкоспециализированных подходов, направленных на устранение частных проблем. Настоящая работа продолжает эту линию, предлагая комплекс модификаций, которые могут быть интегрированы с существующими подходами для дальнейшего повышения эффективности квантования MoE-моделей.

5. Предлагаемый метод

Опираясь на результаты анализа существующих решений и выявленные в них ограничения, в настоящем разделе представлен разработанный автором комплекс взаимосвязанных модификаций, направленных на повышение эффективности посттренировочного квантования MoE-моделей.

5.1 Функция потерь с учетом коэффициентов маршрутизации

В качестве посттренировочного метода квантования в данной работе используется метод GPTQ [18]. Он выполняет квантование весов независимо для каждого линейного слоя сети. Пусть W - веса линейного слоя, а X - входные данные этого слоя, полученные на небольшом подмножестве калибровочных данных. Тогда задача восстановления определяется следующим образом:

$$\underset{w}{\operatorname{argmin}} \cdot \|WX^T - \hat{W}X^T\|_2^2$$

Данная целевая функция, представляющая собой сумму квадратов ошибок, образует квадратичную задачу, что позволяет вычислять жадное оптимальное обновление весов поэлементно с использованием гессиана H . В GPTQ гессиан вычисляется как эмпирическая матрица Фишера по формуле $H = 2XX^T$. Важно, что вычисляется не точный гессиан функции потерь, а его аппроксимация через статистику входов, которую можно эффективно получить за один проход по калибровочным данным. Интуиция подобного вычисления кроется в том, что именно гессиан кодирует информацию о том, как сильно ошибка квантования того или иного веса повлияет на выход слоя: направления с большими значениями в H более чувствительны к ошибке, поэтому метод квантует веса в порядке убывания диагональных элементов H и компенсирует возникшую ошибку на оставшихся, ещё не квантованных весах, минимизируя взвешенную гессианом ошибку.

Стоит отметить, что в данном подходе никак не учитывается структура МоЕ-модели, поскольку каждый линейный слой получает свой набор активаций. В связи с этим в работе предлагается дополнить вычисление гесса за счет коэффициентов маршрутизации:

$$H = 2(X \cdot c)X^T,$$

где c - вектор, состоящий из значений маршрутизатора G для каждого токена из входной последовательности, который будет обрабатываться соответствующим экспертом.

5.2 Сбалансированный набор калибровочных данных

Стандартная процедура посттренировочного квантования языковых моделей с использованием метода GPTQ опирается на калибровочные наборы, сформированные из общедоступных текстовых корпусов, таких как wikitext2 [27] или C4 [28]. Для обычных трансформер-моделей такой подход является оправданным, поскольку каждый параметр модели задействуется при обработке любого входного токена, и статистические свойства активаций, накопленные на достаточно представительной выборке, адекватно отражают чувствительность весов к квантованию.

Однако в случае МоЕ-моделей прямое применение тех же калибровочных наборов приводит к заметному снижению итогового качества квантованной модели. Причина заключается в фундаментальном свойстве МоЕ-слоёв: маршрутизатор направляет различные токены к различным подмножествам экспертов, причём распределение активаций между экспертами может являться крайне неравномерным.

Для количественной оценки степени сбалансированности активаций экспертов в пределах одного МоЕ-блока предлагается использовать энтропийный критерий. Пусть для заданного МоЕ-слоя с номером l , содержащего P экспертов, на некотором калибровочном наборе данных подсчитано число $n_{l,j}$ токенов, которые обрабатываются экспертом с индексом $j \in 1, \dots, P$. Тогда можно определить эмпирическую вероятность активации эксперта j при помощи нормализации:

$$p_{l,j} = \frac{n_{l,j}}{\sum_{i=1}^P n_{l,i}}$$

На основе полученного распределения вероятностей вычисляется энтропия активаций для МоЕ-блока по формуле Шеннона:

$$I_l = -\sum_{i=1}^P p_{l,i} \log p_{l,i}$$

Данная величина достигает максимума $\log P$ в случае строго равномерного распределения активаций между всеми экспертами и стремится к нулю, когда подавляющее большинство токенов направляется к одному или нескольким доминирующим экспертам. Таким образом, энтропия I_l служит удобной скалярной метрикой, характеризующей, насколько равномерно задействованы эксперты внутри конкретного МоЕ-слоя при обработке заданного калибровочного набора.

Для конкретной модели Qwen3-30B-A3B, содержащей 48 МоЕ-блоков по 128 экспертов в каждом, результат вычисления энтропии для разных калибровочных данных представлен на рис. 1. Можно заметить, что для наборов данных wikitext2 [27], C4 [28] и GSM8k [29] среднее значение энтропии примерно на 10-15% ниже максимально возможного, а дисперсия по МоЕ-блокам растет в зависимости от типа домена калибровочных данных. Наиболее контрастная картина наблюдается для GSM8K: в ряде МоЕ-блоков энтропия резко падает, указывая на наличие явно доминирующих экспертов, через которые проходит подавляющая доля токенов, в то время как остальные эксперты активируются эпизодически или не используются вовсе.

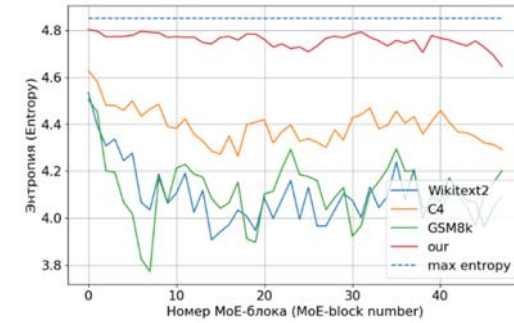


Рис. 1. Зависимость энтропии МоЕ-блоков модели от калибровочных данных для Qwen3-30B-A3B.

Fig. 1. MoE-block entropy for different calibration data for Qwen3-30B-A3B.

Наблюдаемая зависимость энтропии активаций от доменной принадлежности и стилистики калибровочных данных указывает на то, что ни один отдельно взятый набор не способен обеспечить равномерную активацию всех экспертов во всех МоЕ-блоках. В связи с этим в настоящей работе предлагается формировать калибровочный набор путём целенаправленного семплирования последовательностей из нескольких источников, охватывающих различные предметные области (домены) и, возможно, различные естественные языки. Целевой функцией такого семплирования является максимизация минимальной по всем МоЕ-блокам энтропии активаций экспертов.

5.2.1 Алгоритм семплирования калибровочных данных

Ниже представлен формализованный жадный алгоритм построения калибровочного набора, ориентированный на повышение энтропии МоЕ-блоков. Алгоритм предполагает наличие набора исходных корпусов, соответствующих различным доменам и языкам, а также целевой модели, для которой может быть эффективно подсчитана статистика активаций экспертов.

Вход: предобученная МоЕ модель, коллекция наборов данных $\{D_m\}_{m=1}^M$, требуемый размер калибровочного набора B

Выход: калибровочный набор D_{cal}

Шаги алгоритма:

1. Для каждого блока l и каждого эксперта инициализировать счетчики $n_{l,j}$ нулями, задать пустой D_{cal}
2. Повторять до тех пор, пока общий размер D_{cal} не достигнет B :
 - a. Равновероятно выбрать один из наборов данных D_m , извлечь случайную подвыборку S . Если набор исчерпан, допускается его повторное использование с предварительным перемешиванием.
 - b. Пропустить S через модель и обновить счетчики $n_{l,j}$.
 - c. На основе обновлённых счётчиков вычислить текущую энтропию активаций I_l .
 - d. Определить блок с минимальной энтропией: $l^* = \arg \min_l I_l$.
 - e. **Критерий принятия подвыборки:** если энтропия I_{l^*} увеличилась по сравнению с предыдущим значением, зафиксированным для этого же блока, то добавить S в D_{cal} . В противном случае отклонить подвыборку.

Результат вычисления энтропии на модели Qwen3-30-A3B для полученного с помощью данного алгоритма калибровочного набора данных можно наблюдать на рис. 1 в виде кривой с названием «our».

Практические детали по реализации:

- Шаги 1-2 повторялись для определения разных калибровочных данных нужного размера. Выбирался набор с максимальной минимальной энтропией.
- Формирование исходной коллекции наборов данных $\{D_m\}_{m=1}^M$ осуществлялось на основе технического отчета для модели Qwen3-30-A3B. Исходный предобученный корпус является закрытым, однако представленное в отчете описание позволило подобрать из открытых источников наборы данных, воспроизводящие ключевые домены. Их использование поспособствовало получению калибровочной выборки с высокой энтропией относительно базовых наборов.

Таким образом, формируется калибровочный набор, обеспечивающий более равномерную активацию экспертов по сравнению с любым из базовых наборов по отдельности, что, в свою очередь, должно положительно сказаться на качестве последующего квантования методом GPTQ.

5.3 Квантование экспертов в режиме смешанной точности

Используемый в работе метод GPTQ в комбинации с INT форматом квантования по умолчанию оперирует единой разрядностью (INT4, INT3 или INT2) для всех линейных слоев модели. Такой подход, безусловно, удобен с инженерной точки зрения, однако различные параметры, равно как и отдельные эксперты внутри одного МоЕ-блока, демонстрируют неодинаковую чувствительность к понижению точности представления весов. В результате появилось целое направление по квантованию весов с использованием смешанной точности (mixed-precision quantization), которое позволяет назначать различную разрядность разным компонентам модели, как правило целым линейным слоям или определенным частям матриц параметров.

В контексте МоЕ-моделей необходимость применения смешанной точности становится очевидной при анализе текущих результатов современных работ по квантованию. Несколько исследований [23-24] отмечают, что квантование в 4 и даже в 3 бита для МоЕ-архитектур осуществляется практически без потерь по сравнению с FP16. Однако при использовании 2 бит для всех экспертов наблюдается резкая, зачастую катастрофическая просадка в качестве. Такое положение дел подтверждает применение метода GPTQ к моделям Qwen3-30B-A3B, Mixtral-8x7b с использованием фиксированного формата квантования. Результаты представлены на рис. 2. Качество генерации текста, оцениваемое с помощью точности ответа модели на определенные вопросы, при использовании INT2 ухудшается на 25–50% в зависимости от задачи.

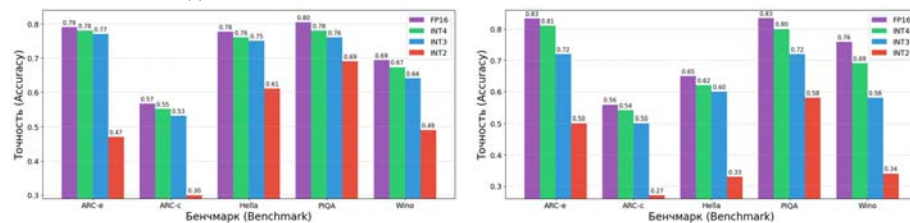


Рис. 2. Метрики качества модели Qwen3-30B-A3B (слева) и Mixtral-8x7b (справа) при различных режимах квантования.

Fig. 2. Zero-shot accuracy for Qwen3-30B-A3B (left) and Mixtral-8x7b (right) with different quantization formats.

Таким образом, оптимальным направлением для текущего сжатия МоЕ-моделей является использование квантования в смешанной битности, а именно, распределение разрядности между INT3 и INT2 для различных экспертов. Ключевой задачей при этом становится разработка метрики, которая позволила бы адекватно определить чувствительность каждого

эксперта к понижению точности и на основе этой оценки принять решение о назначении ему той или иной битности.

В работе предлагается оценка, опирающаяся на результаты, полученные в области прореживания МоЕ-архитектур. В ряде исследований [26] экспериментально установлено, что наименее значимыми с точки зрения финального качества модели оказываются те эксперты, чей вклад в результирующий тензор на выходе МоЕ-блока является минимальным. Формально для эксперта $f_{l,j}$ внутри блока с индексом l вычисление производится следующим образом:

$$S_{l,j} = \frac{1}{|X_{l,j}|} \sum_{x \in X_{l,j}} c_{l,j} \|f_{l,j}(x)\|_2$$

В формуле $X_{l,j}$ - множество токенов для эксперта $f_{l,j}$, которые обрабатываются данным экспертом, что означает $|X_{l,j}| = n_{l,j}$.

Иными словами, мотивация использования данного значения такова, что если для некоторого эксперта произведение нормы его выходного вектора на соответствующий весовой коэффициент маршрутизатора невелико, то более агрессивное квантование такого эксперта с высокой вероятностью не приведёт к существенной деградации качества.

Важное замечание: для моделей под экспертом понимается MLP блок состоящий из комбинации трех линейных слоев. Значение $S_{l,j}$ вычисляется для эксперта целиком. Таким образом, использование данной метрики предполагает, что линейные слои внутри одного эксперта всегда имеют одинаковую битность.

6. Результаты

В данном разделе представлены результаты, полученные в ходе экспериментальной проверки предложенного комплекса модификаций методов квантования для МоЕ-моделей, включая оценку качества на стандартных тестовых задачах, анализ влияния отдельных компонентов предлагаемого подхода и сравнение с существующими решениями в данной области.

Целевая модель: эксперименты проведены для моделей Qwen3-30B-A3B и Mixtral-8x7b [11].

Метод квантования: в данной работе предлагается использовать GPTQ [18] в качестве основного метода квантования.

Конфигурация: квантованию подвергаются в INT3 и INT2 форматы только линейные слои из МоЕ-блоков, остальные параметры остаются в формате FP16.

Калибровочные данные: формирование выборок осуществлялось с использованием наборов данных Wikitext [27], C4 [28], GSM8k [29] (train), OpenWebMath [30], CulturaX [31], Nemotron [32].

Критерии качества: оценка качества генерации осуществлялась с использованием специализированных тестовых наборов данных ARC [33], HellaSwag [34], PIQA [35], Winogrande [36] в 0-shot режиме (в таблицах указывается средняя точность по всем задачам в колонке с названием «0-shot tasks»), а также на MMLU [37], MMLU-Pro [38] и GSM8k [29] в 5-shot режиме, в которых вычисляется доля правильных ответов на текстовые запросы.

В табл.1 приведены результаты (в процентах) квантования целевой модели в INT3 формат (без смешанной точности) для базового варианта GPTQ и для предложенного в работе усовершенствования с домножением гесса на коэффициенты маршрутизатора (в таблице обозначен как «+ routing»). Для каждого из методов вычислены значения для трех различных калибровочных данных: wikitext, GSM8k (train часть) и Sampled (калибровочные данные, собранные при помощи алгоритма из пункта 5.2.1).

Анализ результатов показывает, что модифицированное вычисление гесса приводит к статистически значимому улучшению качества на большинстве 0-shot задач. Иными словами, для тестовых наборов ARC, HellaSwag, PIQA и WinoGrande наблюдается устойчивый

прирост точности, достигающий в отдельных случаях нескольких процентных пунктов. В то же время для более сложных тестовых наборов MMLU и MMLU-Pro изменения оказываются существенно менее выраженными и, как правило, находятся в пределах статистической погрешности. Данный эффект, по всей видимости, объясняется фундаментальными различиями в природе указанных тестовых наборов данных. Тестовые критерии MMLU и MMLU-Pro состоят из вопросов, требующих глубоких предметных знаний из специализированных областей, и являются значительно более трудными для моделей, чем стандартные тесты на здравый смысл. В условиях высокой сложности задачи и, как следствие, большей неопределённости в предсказаниях, локальные улучшения, вносимые послойной оптимизацией квантования, оказывают меньшее влияние на итоговый результат. Таким образом, выигрыш от более точной аппроксимации гессаана нивелируется общей высокой вариативностью ответов модели на специализированных вопросах, что делает прирост качества менее заметным по сравнению с более простыми 0-shot задачами.

Табл. 1. Результаты квантования Qwen3-30B-A3B в INT3 формат с/без использования routing для разных калибровочных данных методом GPTQ. Значения указаны в процентах.
Table 1. Results for INT3 Qwen3-30B-A3B model GPTQ quantization for different calibration data and routing strategy. Values are given in percent.

Метод	Калибровочные данные	0-shot tasks	MMLU	MMLU-pro	GSM8k
FP16	-	72.6	79.6	69.5	87.3
GPTQ	wikitext	66.2	63.4	58.1	75.1
GPTQ + routing		66.7	63.2	57.8	78.1
GPTQ	GSM8k	63.2	60.1	57.3	84.2
GPTQ + routing		63.1	60.4	58.3	84.3
GPTQ	Sampled	66.5	63.5	60.2	84.1
GPTQ + routing		67.2	64.1	60.1	84.5

Сравнение результатов для различных калибровочных наборов позволяет сделать следующие наблюдения. Наилучшие показатели качества достигаются при использовании предложенного в работе набора данных (обозначенного как «Sampled»), что согласуется с выдвинутым ранее предположением о необходимости максимизации энтропии активаций экспертов для повышения эффективности квантования. Примечательно, что за счёт применения предложенной процедуры формирования сбалансированной выборки удалось добиться улучшения метрик на всём спектре рассмотренных тестовых критериев по сравнению со стандартным калибровочным набором Wikitext. Отдельного внимания заслуживает случай использования набора данных GSM8K в качестве калибровочного. Как видно из табл. 1, такая калибровка приводит к выраженной деградации качества на подавляющем большинстве тестовых наборах данных, в то время как на GSM8k точность сохраняется на сравнительно высоком уровне. Подобное поведение с высокой долей вероятности объясняется узкой доменной специализацией калибровочных данных: GSM8k содержит исключительно математические задачи, что формирует смещённую статистику активаций экспертов (и, соответственно, низкую энтропию, рис. 1).

Для оценки практической применимости метрики чувствительности экспертов была проведена серия экспериментов по квантованию моделей в формате смешанной точности. Использовались все предложенные ранее модификации для метода GPTQ. В соответствии с описанной ранее процедурой, для каждого эксперта вычислялось значение $S_{l,j}$, после чего

эксперты ранжировались в порядке возрастания данной величины. Наименее чувствительным экспертам (с наименьшими значениями $S_{l,j}$) назначалась разрядность 2 бита, тогда как остальные эксперты квантовались в 3 бита. Варьирование доли экспертов, переводимых в 2-битный формат, позволило получить набор конфигураций с различной средней битностью.

На рис. 3 представлена зависимость точности квантованных моделей на тестовых наборах GSM8K и MMLU от средней битности экспертов, а также отмечен уровень качества, соответствующий базовому квантованию в INT3. Можно заметить, что снижение разрядности для 10% наименее чувствительных экспертов (что соответствует снижению средней битности до 2.9) приводит к пренебрежимо малому падению точности. При увеличении доли 2-битных экспертов до 20-30% (средняя битность 2.7-2.8) потери остаются умеренными и могут считаться приемлемыми в сценариях, где приоритет отдаётся уменьшению размера модели, а не сохранению абсолютной точности.

В табл. 2 приведено сравнение предложенного метода с существующими подходами. Следует отметить, что для большинства методов в открытых источниках отсутствуют опубликованные результаты применительно к модели Qwen3-30B-A3B, поэтому сопоставление ограничено подходами RTN, GPTQ и MoEQuant.

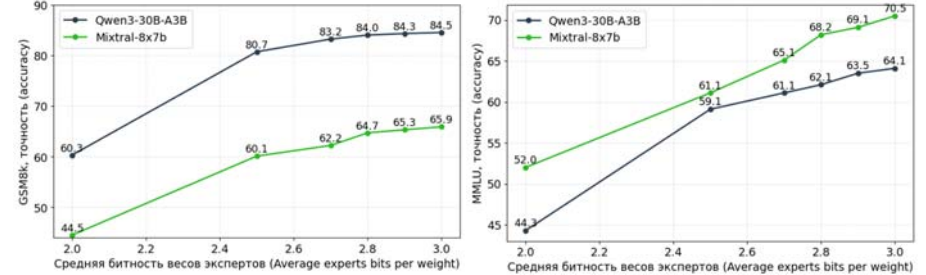


Рис. 3. Изменение качества на тестовых наборах GSM8k (слева) и MMLU (справа) при использовании смешанной точности.
Fig. 3. GSM8k (left) and MMLU (right) accuracy for different average bits per weight quantization settings.

Табл. 2. Сравнение метода GPTQ с предложенными модификациями против существующих алгоритмов по квантованию MoE для модели Qwen3-30B-A3B.
Table 2. Comparison between GPTQ with proposed modifications and existed solutions for Qwen3-30B-A3B.

Метод	Средняя битность	0-shot tasks	MMLU	GSM8k
FP16	16	72.6	79.6	87.3
RTN		~ 25.0	~ 20.0	~ 25.0
GPTQ	3	66.2	63.4	75.1
MoeQuant		66.8	64.1	83.8
GPTQ +routing +sampled data		67.2	64.1	84.5
GPTQ +routing +sampled data +mixed precision	2.8	66.7	62.1	84.0

В конфигурации INT3 предложенный метод демонстрирует наилучшие показатели среди всех рассмотренных, превосходя как стандартный GPTQ, так и MoEQuant. При переходе к смешанной точности со средней битностью 2.8 качество модели ожидаемо несколько снижается относительно варианта INT3, однако остаётся сопоставимым с результатами MoEQuant. Таким образом, предложенный подход позволяет достичь более высокой степени сжатия при сохранении конкурентоспособного уровня точности.

7. Заключение

В настоящей работе рассмотрена задача повышения эффективности посттренировочного квантования больших языковых моделей, построенных по архитектуре «смесь экспертов». Анализ существующих решений свидетельствует, что прямое использование методов квантования, отработанных на стандартных трансформер-моделях (например, GPTQ), в контексте MoE-архитектур сопровождается ощутимой потерей качества. Это обусловлено пренебрежением двумя ключевыми особенностями MoE: разреженной и несбалансированной активацией экспертов, а также взвешивающим влиянием маршрутизатора на их выходы. В рамках исследования предложен комплекс взаимосвязанных модификаций базового алгоритма GPTQ, направленных на преодоление указанных ограничений. Во-первых, разработана процедура формирования калибровочного набора путём семплирования последовательностей из нескольких доменов, ориентированная на максимизацию энтропии активаций экспертов в каждом MoE-блоке. Экспериментально подтверждено, что предложенный подход позволяет улучшить качество квантованной модели на всём спектре рассмотренных тестовых задач. Во-вторых, предложена модификация вычисления гессии, включающая коэффициенты маршрутизатора. Данное изменение позволяет более точно аппроксимировать вклад каждого эксперта в ошибку квантования и, как следствие, эффективнее компенсировать возникающие искажения. В-третьих, на основе анализа работ по прореживанию MoE-моделей введена метрика чувствительности эксперта. Показано, что данная метрика позволяет эффективно ранжировать экспертов по степени допустимой агрессивности квантования, что в результате позволяет уменьшить среднюю битность экспертов на 0.2 без особой потери в качестве генерации текстов.

Список литературы / References

- [1]. Achiam J. et al. Gpt-4 technical report. Open AI. arXiv preprint, 2023. DOI: 10.48550/arXiv.2303.08774.
- [2]. Liu A. et al. Deepseek-v3 technical report. arXiv preprint, 2024. DOI: 10.48550/arXiv.2412.19437.
- [3]. Aminabadi R.Ya. et al. DeepSpeed-inference: enabling efficient inference of transformer models at unprecedented scale. SC22: International Conference for High Performance Computing, Networking, Storage and Analysis. IEEE, 2022.
- [4]. Lin J. et al. AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration. Proceedings of machine learning and systems 6, 2024, pp. 87-100.
- [5]. Masoudnia S., Ebrahimpour R. Mixture of experts: a literature survey. Artificial Intelligence Review 42.2, 2014, pp. 275-293.
- [6]. Yuksel S.E., Wilson J.N., Gader P.D. Twenty years of mixture of experts. IEEE transactions on neural networks and learning systems. 2012, vol. 23, no. 8, pp. 1177-1193.
- [7]. Jacobs R.A. et al. Adaptive mixtures of local experts. Neural computation, vol. 3, 1991, pp. 79-87.
- [8]. Lepikhin D. et al. Gshard: Scaling giant models with conditional computation and automatic sharding. arXiv preprint, 2020. DOI: 10.48550/arXiv.2006.16668.
- [9]. Fedus W., Zoph B., Shazeer N. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. Journal of Machine Learning Research, 2022, vol. 23, pp. 120:1-120:39.
- [10]. Vaswani A. et al. Attention is all you need. Advances in neural information processing systems 30, 2017.
- [11]. Yang A. et al. Qwen3 technical report. arXiv preprint, 2025. DOI: 10.48550/arXiv.2505.09388.
- [12]. Gou J. et al. Knowledge distillation: A survey. International journal of computer vision, 2021, vol. 129, pp. 1789-1819. DOI: 10.1007/s11263-021-01453-z.

- [13]. Stanton S. et al. Does knowledge distillation really work? Advances in neural information processing systems, 2021, vol. 34, pp. 6906-6919.
- [14]. Ma X., Fang G., Wang X. Llm-pruner: On the structural pruning of large language models. Advances in neural information processing systems, 2023, vol. 36, pp. 21702-21720. DOI: 10.52202/075280-0731.
- [15]. Cheng H., Zhang M., Shi J.Q. A survey on deep neural network pruning: Taxonomy, comparison, analysis, and recommendations. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2024, vol. 46, no. 12, pp. 10558-10578.
- [16]. Dettmers T. et al. Qlora: Efficient finetuning of quantized llms. Advances in neural information processing systems, 2023, vol. 36, pp. 10088-10115.
- [17]. Han Z. et al. Parameter-efficient fine-tuning for large models: A comprehensive survey. arXiv preprint, 2024. DOI: 10.48550/arXiv.2403.14608.
- [18]. Frantar E. et al. GPTQ: Accurate post-training quantization for generative pre-trained transformers. arXiv preprint, 2022. DOI: 10.48550/arXiv.2210.17323.
- [19]. Liu Z. et al. LLM-QAT: Data-free quantization aware training for large language models. Findings of the Association for Computational Linguistics: ACL 2024, 2024.
- [20]. Malinovsky V. et al. PV-tuning: Beyond straight-through estimation for extreme llm compression. Advances in Neural Information Processing Systems, 2024, vol. 37, pp. 5074-5121.
- [21]. Li Yu. et al. GPTAG: Efficient finetuning-free quantization for asymmetric calibration. arXiv preprint, 2025. DOI: 10.48550/arXiv.2504.02692 (2025).
- [22]. Touvron H. et al. LLaMA 2: Open foundation and fine-tuned chat models. arXiv preprint, 2023. DOI: 10.48550/arXiv.2307.09288.
- [23]. Li P. et al. QuantMoE-Bench: Examining Post-Training Quantization for Mixture-of-Experts. arXiv preprint, 2024. DOI: 10.48550/arXiv.2406.08155.
- [24]. Hu X. et al. Moequant: Enhancing quantization for mixture-of-experts large language models via expert-balanced sampling and affinity guidance. arXiv preprint, 2025. DOI: 10.48550/arXiv.2505.03804.
- [25]. Huang W. et al. Mixture compressor for mixture-of-experts llms gains more. arXiv preprint, 2024. DOI: 10.48550/arXiv.2410.06270.
- [26]. Liu J. et al. A survey on inference optimization techniques for mixture of experts models. ACM Computing Surveys, 2026, vol. 58, no. 10, pp. 1-37.
- [27]. Johnson I., Kaffee L.-A., Redi M. Wikimedia data for AI: a review of Wikimedia datasets for NLP tasks and AI-assisted editing. Proceedings of the First Workshop on Advancing Natural Language Processing for Wikipedia, 2024.
- [28]. Colin R. Exploring the limits of transfer learning with a unified text-to-text transformer. J. Mach. Learn. Res., 2020, vol. 21.
- [29]. Plaat A. et al. Multi-step reasoning with large language models, a survey. ACM Computing Surveys, 2025, vol. 58, no. 6, pp. 1-35.
- [30]. Paster K. et al. Openwebmath: An open dataset of high-quality mathematical web text. arXiv preprint, 2023. DOI: 10.48550/arXiv.2310.06786 (2023).
- [31]. Thuat Nguyen et al. CulturaX: A Cleaned, Enormous, and Multilingual Dataset for Large Language Models in 167 Languages. LREC-COLING 2024. DOI: 10.48550/arXiv.2309.09400.
- [32]. Blakeman A. NVIDIA Nemotron 3: Efficient and Open Intelligence. arXiv preprint, 2025. DOI: 10.48550/arXiv.2512.20856.
- [33]. Clark P. Think you have solved question answering? Try arc, the ai2 reasoning challenge. arXiv preprint, 2018. DOI: 10.48550/arXiv.1803.05457.
- [34]. Zellers R. Hellaswag: Can a machine really finish your sentence? Proceedings of the 57th annual meeting of the association for computational linguistics. 2019.
- [35]. Bisk Y. Piqa: Reasoning about physical commonsense in natural language. Proceedings of the AAAI conference on artificial intelligence, 2020, vol. 34, no. 5.
- [36]. Sakaguchi K. Winogrande: An adversarial winograd schema challenge at scale. Communications of the ACM, 2021, vol. 64, no. 9, pp. 99-106.
- [37]. Hendrycks D. et al. Measuring massive multitask language understanding. arXiv preprint, 2020. DOI: 10.48550/arXiv.2009.03300 (2020).
- [38]. Wang Yu. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. Advances in Neural Information Processing Systems, 2024, vol. 37, pp. 95266-95290.

Информация об авторах / Information about authors

Дмитрий Романович ПОЙМАНОВ – окончил магистратуру кафедры математических методов прогнозирования (ММП) факультета вычислительной математики и кибернетики (ВМК) Московского государственного университета имени М. В. Ломоносова в 2024 году. В настоящее время является аспирантом ММП ВМК МГУ. Сфера научных интересов: искусственный интеллект, большие языковые модели, сжатие нейросетей и оптимизация вычислений.

Dmitrii Romanovich POIMANOV – graduated from the master’s program of the Department of the Mathematical Methods of Forecasting (MMP) at the Faculty of the Computational Mathematics and Cybernetics (CMC) of Lomonosov Moscow State University in 2024. He is currently a graduate student at MSU. Areas of scientific interests: artificial intelligence, large language models, neural network compression and optimization.