Раздел 3. «IT-технологии, энергетика, автоматизация и вычислительная техника»

МРНТИ 27.35.15
УДК 004.8
DOI 10.53002/012

И.Р. Дашкин, Г.Д. Когай

Карагандинский технический университет имени Абылкаса Сагинова, Караганда, Казахстан
(E-mail: idashkin00@mail.ru, g.kogay@mail.ru)

Малые языковые модели: компромисс между эффективностью и производительностью в эпоху больших языковых моделей

В данной статье исследуется компромисс между эффективностью и производительностью малых языковых моделей (SLM) с менее чем 15 миллиардами параметров, что представляет собой актуальную альтернативу ресурсоемким большим языковым моделям (LLM). Проведено сравнение современных SLM, таких как Llama 3.2 3B и Llama 3 8B, Gemma 2B и 9B, Qwen2.5 3B, 7B и 14B, Phi-4 14B и Mistral NeMo 12B, с использованием стандартизированных бенчмарков (MMLU-PRO, GPQA, IFEval, MATH, BBH) для оценки их способностей в генерации текста, суммаризации, ответах на вопросы и логических рассуждениях. Результаты показывают, что некоторые SLM демонстрируют производительность, близкую к высокопараметрическим моделям, таким как GPT-4o, при значительно меньших вычислительных затратах. Работа подчеркивает потенциал SLM для создания более доступных и экологичных решений в области искусственного интеллекта, предлагая практические рекомендации для исследователей и разработчиков.

Ключевые слова: Малые языковые модели, модели с малым числом параметров, обработка естественного языка, сравнение языковых моделей, Llama 3.2 3B и Llama 3 8B, Gemma 2B и 9B, Qwen2.5 3B 7B и 14B, Phi-4 14B, Mistral NeMo 12B, энергоэффективность, производительность.

Введение

В последние годы в области обработки естественного языка (NLP) наблюдаются значительные успехи в разработке больших языковых моделей (LLM). Эти модели, часто состоящие из сотен миллиардов параметров, продемонстрировали беспрецедентные возможности в таких задачах, как генерация текста, перевод и рассуждения. Однако вычислительные затраты, энергопотребление и требования к инфраструктуре, связанные с обучением и развертыванием таких высокопараметрических моделей, вызывают сомнения в их практичности и устойчивости. В результате растет интерес к исследованию более компактных, более эффективных малых языковых моделей (SLM), которые могут обеспечить конкурентоспособную производительность, будучи при этом более доступными и экологичными.

В данной статье исследуется компромисс между эффективностью и производительностью в малопараметрических SLM, фокусируясь на моделях с менее чем 15 миллиардами параметров. В частности, мы сравниваем производительность нескольких современных малопараметрических SLM, включая Llama 3.2 3B и Llama 3 8B (разработанная Meta) [1], Gemma 2B и Gemma 9B (от Google) [2], Qwen2.5 3B, 7B и 14B (от Alibaba) [3], Phi-4 14B (от Microsoft) [4]. Mistral NeMo 12B (от MistralAI) [5]. Эти модели представляют собой последние достижения в области компактного языкового моделирования, используя инновационные архитектуры и методы обучения для достижения высокой производительности при значительно меньшем количестве параметров.

Чтобы контекстуализировать производительность этих небольших моделей, также будет включен GPT-4o в качестве опорной точки. Хотя GPT-4o является высокопараметрической моделью, её включение позволяет лучше понять разрыв в производительности и компромиссы между малыми и крупными моделями. Оценивая эти модели на ряде задач, таких как генерация текста, суммаризация, ответы на вопросы и логические рассуждения, мы стремимся определить, какие малые языковые модели предлагают наилучший баланс между эффективностью и производительностью.

Раздел 3. «IT-технологии, энергетика, автоматизация и вычислительная техника»

Результаты этого исследования имеют важное значение как для исследователей, так и для практиков. Для исследователей эта работа подчеркивает потенциал малых языковых моделей в достижении производительности, близкой к современным стандартам, при сниженных вычислительных затратах. Для практиков она предоставляет полезные рекомендации по выбору моделей, наиболее подходящих для конкретных приложений, особенно в условиях ограниченных ресурсов. В конечном итоге данная статья вносит вклад в продолжающуюся дискуссию о том, как сделать искусственный интеллект более эффективным, доступным и устойчивым.

Методы и материалы

Чтобы всесторонне оценить производительность малопараметрических SLM, был выбран набор стандартизированных бенчмарков, которые проверяют различные аспекты понимания языка, рассуждений и решения задач. В вычислительной технике бенчмарк - это выполнение компьютерной программы, набора программ или других операций для оценки относительной производительности объекта, обычно путем проведения ряда стандартных тестов и испытаний. Эти бенчмарки широко известны в сообществе NLP и обеспечивают надежную основу для сравнения возможностей моделей. В число выбранных бенчмарков входят:

1. MMLU-PRO – Massive Multitask Language Understanding - Progressive (Массовое многозадачное понимание языка – Прогрессивное)

Измеряет способность модели выполнять многозадачное обучение в различных областях, включая гуманитарные, технические и социальные науки. Этот критерий оценивает обобщаемость и широту знаний в LLM с малыми параметрами, гарантируя, что они смогут справиться с широким спектром задач без тонкой настройки под конкретную область [6].

2. GPQA – General Purpose Question Answering (Ответы на вопросы общего назначения)

Проверяет способность модели отвечать на сложные, открытые вопросы, требующие глубоких рассуждений и синтеза информации. GPQA особенно полезен для оценки способности к рассуждениям малопараметрических LLM, поскольку он заставляет модели выходить за рамки поверхностного понимания [7].

3. IFEval – Instruction Following Evaluation (Оценка следования инструкций)

Оценивает, насколько хорошо модель может следовать явным и неявным инструкциям в задачах генерации текста. Этот критерий важен для оценки практической пригодности малопараметрических LLM в реальных приложениях, где следование инструкциям имеет первостепенное значение [8].

4. MATH (Решение математических задач)

Оценивает способность модели решать математические задачи, начиная с базовой арифметики и заканчивая сложными вычислениями. MATH дает представление о логических и аналитических возможностях малопараметрических SLM, которые важны для приложений в образовании, финансах и научных исследованиях [9].

5. BBH – BIG-Bench Hard

BBH включает задачи, которые особенно сложны для небольших моделей, такие как причинно-следственные рассуждения, аналоговые рассуждения и разбор смысла слов. BBH разработан, чтобы расширить границы возможностей малопараметрических SLM, что делает его идеальным бенчмарком для выявления их сильных и слабых сторон в сложных задачах рассуждения [10].

Результаты и обсуждение

1. Бенчмарк MMLU-PRO

На рисунке 1 показана производительность нескольких малопараметрических языковых моделей в бенчмарке MMLU-PRO с использованием GPT-4o в качестве высокопараметрической эталонной модели. MMLU-PRO (Massive Multitask Language Understanding - Professional) оценивает модели на широком спектре профессиональных задач, проверяя их способность к обобщению и практическую полезность в различных областях. Вот как выглядят эти модели:

Раздел 3. «IT-технологии, энергетика, автоматизация и вычислительная техника»

GPT-4o лидирует с результатом 73 балла, что отражает его превосходную производительность во всех эталонных задачах. Как и ожидалось, большое количество параметров GPT-4o позволяет ему преуспеть в понимании и рассуждениях, устанавливая сильную точку отсчета для сравнения.

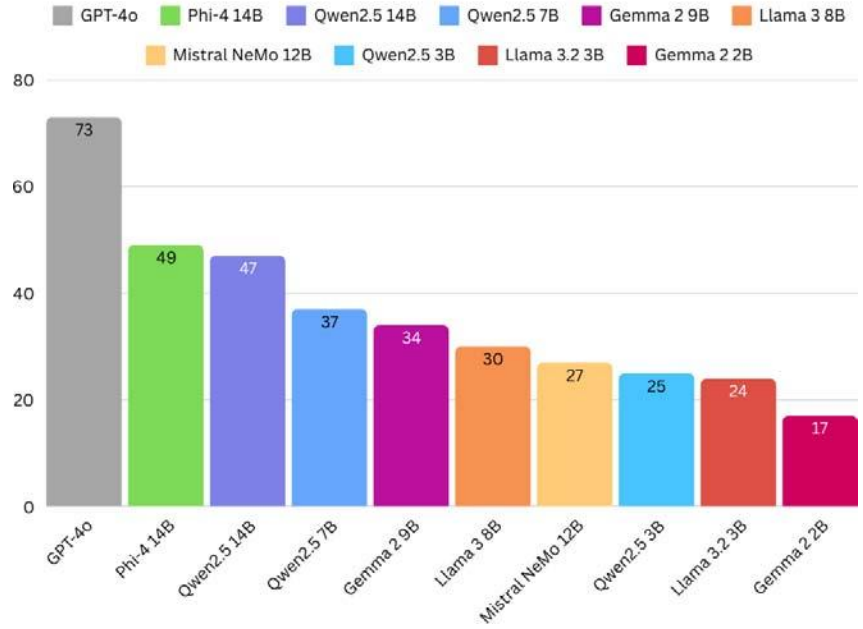

Рисунок 1– Результат бенчмарка MMLU-PRO

Среди моделей с малым числом параметров Phi-4 14B показал наилучший результат, набрав 49 баллов. Это говорит о том, что, несмотря на меньшее количество параметров по сравнению с GPT-4o, Phi-4 14B показывает высокие результаты, особенно в профессиональных задачах, требующих сложных рассуждений и навыков решения проблем.

За ним следует Qwen2.5 14B с результатом 47 баллов, что подтверждает идею о том, что модели с параметрами в диапазоне 10-15 миллиардов могут поддерживать конкурентоспособные результаты в заданиях, требующих понимания языка на высоком уровне.

Qwen2.5 7B набрала 37 баллов, что, хотя и ниже, чем у ее коллеги 14B, все же является высоким результатом, особенно для модели со значительно меньшим количеством параметров.

Такие модели, как Gemma 2 9B (34) и Llama 3 8B (30), показывают хорошие результаты для своего размера, демонстрируя, что небольшие модели могут справляться с задачами профессионального уровня с достаточной степенью точности. Эти модели перспективны для сценариев, где необходимы эффективность и производительность.

Mistral NeMo 12B набирает 27 баллов, что ставит его в средний ряд протестированных моделей. Она демонстрирует сбалансированный подход между производительностью и количеством параметров, что делает ее универсальной для различных приложений.

Модели Qwen2.5 3B (25), Llama 3.2 3B (24) и Gemma 2 2B (17) демонстрируют ожидаемые компромиссы, которые возникают при меньшем бюджете параметров. Хотя эти модели не могут сравниться с моделями с более высокими параметрами по абсолютной производительности, они все же могут достаточно хорошо справляться с менее сложными задачами в рамках MMLU-PRO.

2. Бенчмарк GPQA

На рисунке 2 представлены результаты работы различных языковых моделей с малыми параметрами в эталоне GPQA (General-Purpose Question Answering), при этом в качестве высокопараметрической эталонной модели снова используется GPT-4o. Эталон GPQA оценивает модели по их способности справляться с различными задачами по ответам на вопросы, проверяя их способность к рассуждению, пониманию и знанию фактов в различных областях. Ниже представлен анализ результатов:

GPT-4o явно доминирует в данном бенчмарке, набрав 54 балла, что подтверждает его превосходную способность решать задачи, связанные с ответами на вопросы. Значительный разрыв между GPT-4o и

Раздел 3. «IT-технологии, энергетика, автоматизация и вычислительная техника»

более мелкими моделями подчеркивает преимущества высокопараметрических моделей в достижении современной производительности в задачах GPQA, требующих сложных рассуждений и глубокого понимания фактической информации.

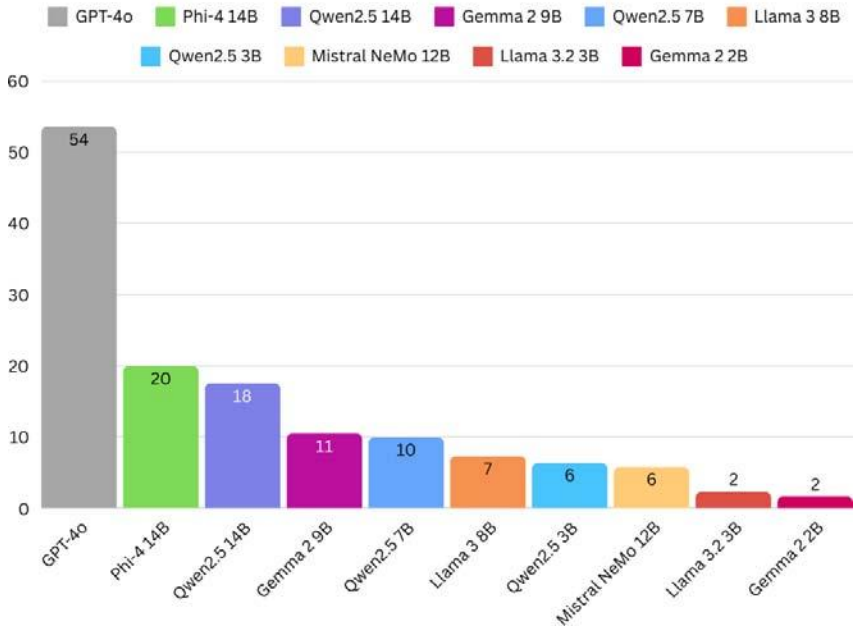

Рисунок 2 – Результат бенчмарка GPQA

Phi-4 14B показывает лучший результат среди малых моделей, набирая 20 баллов. Несмотря на отставание от GPT-4o, Phi-4 14B показывает, что модель со значительно меньшим числом параметров все же может показывать достойные результаты при ответе на вопросы общего назначения. Это делает Phi-4 14B сильным кандидатом для приложений, где крупномасштабные модели вроде GPT-4o слишком требовательны к ресурсам.

Qwen2.5 14B следует вплотную за ней, набрав 18 баллов, что еще больше подтверждает силу моделей в диапазоне 10-15 миллиардов параметров. И Phi-4 14B, и Qwen2.5 14B демонстрируют, что можно достичь относительно высокой производительности в задачах GPQA без вычислительного бремени массивной модели.

Gemma 2 9B набирает 11 баллов, заметно снижая производительность по сравнению с моделями с более высокими параметрами. Однако она все еще демонстрирует достойную компетентность, что делает ее подходящей для сред, где размер модели является ограничивающим фактором.

Qwen2.5 7B набирает 10 баллов, что близко к Gemma 2 9B. Этот результат свидетельствует об уменьшении отдачи от сокращения параметров при решении задач, связанных с ответами на вопросы: дальнейшее уменьшение размера модели приводит к компромиссам в производительности, которые могут оказаться неприемлемыми для более сложных или критически важных приложений.

Такие модели, как Llama 3 8B (7), Qwen2.5 3B (6) и Mistral NeMo 12B (6), занимают середину группы. Эти модели демонстрируют более низкую производительность в задачах GPQA, но все еще могут быть жизнеспособны для более простых или специфических приложений для ответов на вопросы, где эффективность приоритетнее точности.

Llama 3.2 3B и Gemma 2 2B, получившие по 2 балла, являются самыми слабыми в этом бенчмарке. Хотя они демонстрируют ограниченные возможности для ответов на вопросы общего назначения, эти очень маленькие модели все же могут быть полезны для базовых или узкоспециальных приложений в средах с ограниченными ресурсами.

3. Бенчмарк IFEval

Рисунок 3 иллюстрирует производительность различных малопараметрических языковых моделей в эталоне IFEval, снова сравнивая их с GPT-4o в качестве эталона. Эталон IFEval оценивает модели на основе их способности интерпретировать фактическую информацию и отвечать на вопросы,

Раздел 3. «IT-технологии, энергетика, автоматизация и вычислительная техника»

связанные с умозаключениями, что требует сочетания способности к запоминанию знаний и способности к рассуждению. Ниже представлен анализ результатов:

GPT-4o продолжает лидировать с результатом 90 баллов, демонстрируя свою превосходную способность справляться с задачами умозаключения.

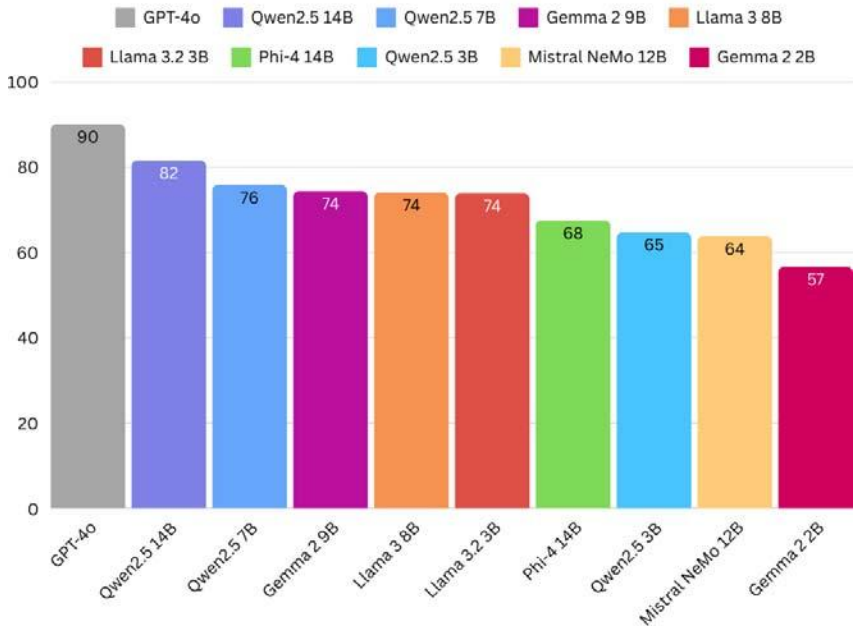

Рисунок 3 – Результат бенчмарка IFEval

Высокий балл подтверждает, что крупномасштабные модели могут хранить и запоминать большие объемы фактических данных, выполняя при этом сложные рассуждения. Это соответствует ожиданиям, поскольку такие модели, как GPT-4o, предназначены для решения широкого спектра задач с большим количеством информации и глубоким пониманием контекста.

Qwen2.5 14B следует вплотную за ней, набрав 82 балла, что вполне конкурентоспособно, учитывая меньший размер параметров по сравнению с GPT-4o. Такая производительность указывает на то, что Qwen2.5 14B хорошо подходит для задач вывода и может быть жизнеспособной альтернативой в средах, где высокопараметрическая модель непрактична из-за ограниченности ресурсов.

Qwen2.5 7B и Gemma 2 9B получили 76 и 74 балла, показывая, что модели с меньшим количеством параметров могут адекватно работать в бенчмарках, основанных на выводах. Разница в производительности между Qwen2.5 14B и Qwen2.5 7B говорит о том, что количество параметров действительно способствует повышению производительности, но не в такой степени, как при уменьшении с 14 до 7B.

Модели Llama 3 8B и Llama 3.2 3B, получившие 74 балла, демонстрируют впечатляющую производительность, несмотря на меньший размер параметров. Модели Llama 3 явно конкурентоспособны, когда речь идет об умозаключениях и фактических рассуждениях, что делает их сильными кандидатами для задач, где важны и производительность, и эффективность.

Phi-4 14B набирает 68 баллов, что ниже, чем ожидалось для модели такого размера. Это может свидетельствовать о том, что, хотя Phi-4 14B превосходит другие бенчмарки, такие как GPQA, его архитектура может быть не так оптимизирована для задач умозаключения, или же это может отражать компромисс в том, как модель балансирует между рассуждениями общего назначения и выполнением конкретных задач.

Qwen2.5 3B (65) и Mistral NeMo 12B (64) довольно близки по производительности, причем Qwen2.5 3B превосходит свои возможности, учитывая меньшее количество параметров. Эти результаты говорят о том, что Mistral NeMo 12B, несмотря на больший размер параметров, не

Раздел 3. «IT-технологии, энергетика, автоматизация и вычислительная техника»

обеспечивает значительного прироста производительности в задачах вывода по сравнению с более компактными и эффективными моделями, такими как Qwen2.5 3B.

Gemma 2 2B, получившая 57 баллов, демонстрирует самую слабую производительность в этом бенчмарке. Небольшой размер модели ограничивает ее способность конкурировать с более крупными моделями в плане вывода фактов, хотя она все еще может быть жизнеспособной в средах, где крайне низкое использование ресурсов является приоритетом по сравнению с точностью задачи.

4. Бенчмарк MATH

Рисунок 4 демонстрирует производительность моделей с малыми параметрами в эталоне MATH, сравнивая их с эталонным GPT-4o. Эталон MATH известен своими сложными задачами по решению математических проблем, которые требуют сильных способностей к рассуждениям, алгоритмического мышления и способности понимать формальные представления математических концепций.

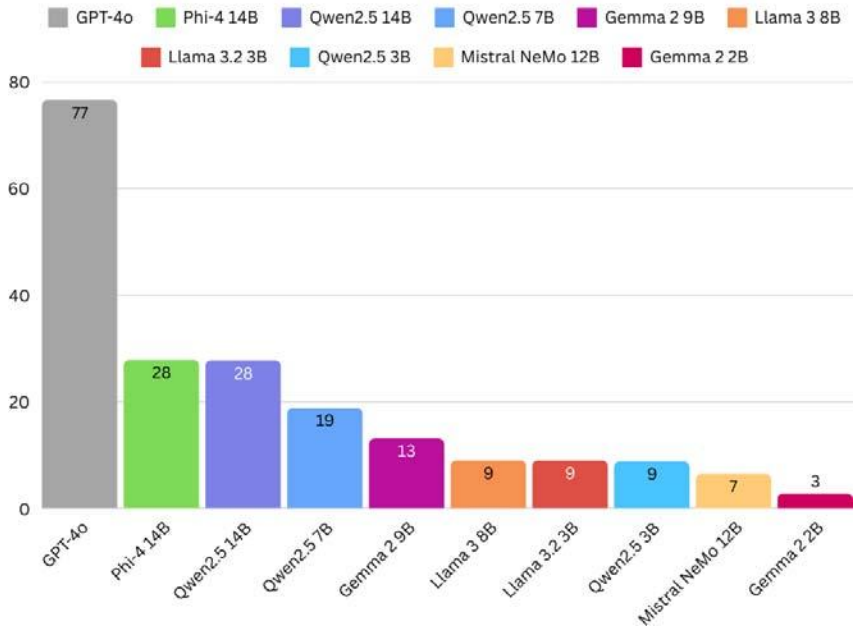

Рисунок 4 – Результат бенчмарка MATH

GPT-4o набрал 77 баллов, значительно опередив другие модели. Этот результат подчеркивает превосходные возможности GPT-4o в решении сложных математических задач, что согласуется с его более широкой, универсальной архитектурой и обширными знаниями, полученными в результате обучения на больших массивах данных.

Phi-4 14B и Qwen2.5 14B, набравшие 28 баллов, являются лучшими моделями меньшего размера. Эти результаты говорят о том, что данные модели обладают хорошими способностями к математическому мышлению, хотя и значительно отстают от GPT-4o. Последовательная производительность Phi-4 14B во многих бенчмарках говорит о том, что это сбалансированная модель, способная решать не только задачи общего назначения, но и в определенной степени специализированные задачи, такие как математика.

Qwen2.5 7B набирает меньшее количество баллов - 19, что говорит о том, что, несмотря на относительную производительность, уменьшение количества параметров приводит к заметному снижению производительности в математических вычислениях. Разница между Qwen2.5 14B и Qwen2.5 7B подчеркивает, что такие задачи, как математика, требующие рассуждений более высокого порядка, значительно выигрывают от дополнительных параметров.

Gemma 2 9B набрала 13 баллов, что говорит о том, что она испытывает больше трудностей с эталоном MATH по сравнению с другими заданиями. Это может говорить о том, что архитектура Gemma 2 менее оптимизирована для формальных и жестких процессов рассуждений, необходимых для решения математических задач, хотя она и показала лучшие результаты в таких задачах, как запоминание фактов (как показано в IFEval).

Раздел 3. «IT-технологии, энергетика, автоматизация и вычислительная техника»

Llama 3 8B, Llama 3.2 3B и Qwen2.5 3B получили по 9 баллов, что свидетельствует о том, что, хотя эти модели компетентны в некоторых других задачах, они сталкиваются со значительными трудностями при работе с математическими рассуждениями на высоком уровне. Количество параметров ограничивает их способность эффективно работать со сложными алгоритмами и уравнениями.

Mistral NeMo 12B набирает 7 баллов, что ниже, чем ожидалось, учитывая относительно большой размер параметров. Это говорит о том, что его обучение или архитектура, возможно, не так ориентированы на математические рассуждения, как у других эталонов, что указывает на компромисс между общей обработкой языка и специфическими возможностями решения задач по математике.

Gemma 2 2B, получившая оценку 3, демонстрирует, что очень маленькие модели испытывают значительные трудности при решении математических задач. Это соответствует ожиданиям, поскольку математические рассуждения требуют более высоких уровней абстракции, символьных манипуляций и логического мышления, с которыми маленькие модели с ограниченными возможностями справляются хуже.

5. Бенчмарк BBH

На рисунке 5 сравниваются модели с малыми параметрами и GPT-4o на эталоне BBH (BigBench Hard), состоящем из множества сложных задач на рассуждение. Эталон BBH известен своими сложными задачами на рассуждение, обобщение и сложные когнитивные задачи, которые требуют более глубокого понимания как абстрактных, так и прикладных концепций в различных областях.

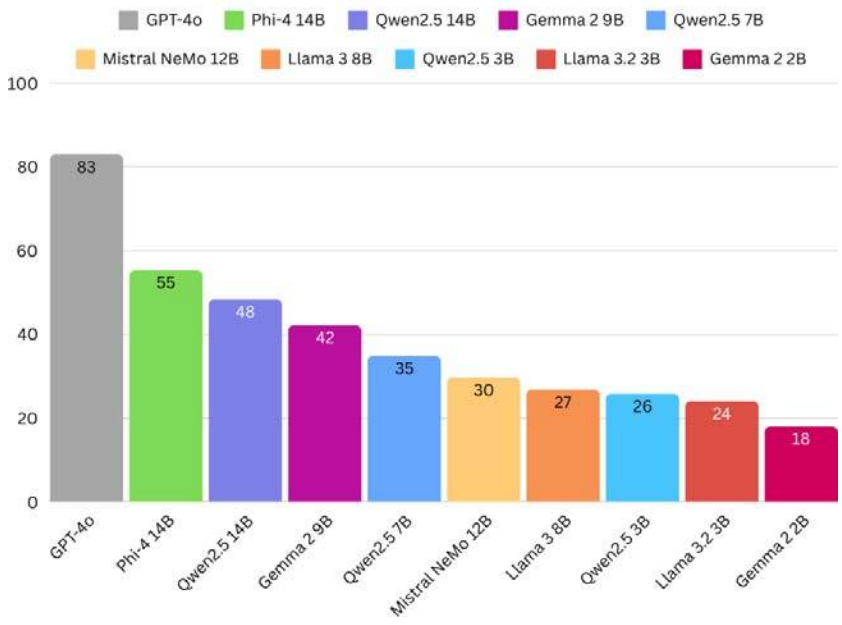

Рисунок 5 – Результат бенчмарка BBH

GPT-4o, набрав 83 балла, остается самой сильной моделью в эталоне BBH с существенным отрывом. Этот результат свидетельствует о превосходных способностях GPT-4o к общему рассуждению, что совпадает с его результатами в других эталонах, таких как IFEval и MATH. Устойчивость модели при решении абстрактных задач отражает ее широкое и эффективное обучение на широком спектре задач и наборов данных.

Phi-4 14B набирает 55 баллов, позиционируя себя как ведущая модель меньшего размера в этом бенчмарке. Высокие результаты во всех бенчмарках, включая BBH, свидетельствуют о том, что Phi-4 14B способна решать более сложные задачи рассуждения, хотя она все еще демонстрирует заметное отставание от GPT-4o. Этот результат указывает на то, что модели с большими параметрами, такие как Phi-4 14B, превосходят по производительности задачи рассуждений, требующие глубоких выводов, абстрагирования и способности к обобщению в различных контекстах.

Раздел 3. «IT-технологии, энергетика, автоматизация и вычислительная техника»

Qwen2.5 14B, набравшая 48 баллов, является следующей по результативности среди небольших моделей. Хотя она не дотягивает до Phi-4 14B, она все равно демонстрирует сильные способности к общему рассуждению, что говорит о том, что ее архитектура хорошо справляется со сложными и абстрактными задачами рассуждения. Разрыв между Phi-4 14B и Qwen2.5 14B позволяет предположить, что специальные архитектурные оптимизации в Phi-4 14B могут дать ему преимущество в решении более абстрактных задач рассуждения.

Gemma 2 9B, получившая 42 балла, демонстрирует компетентность в задачах рассуждения, но все еще отстает от своих более крупных аналогов. Результаты этой модели подчеркивают, что, хотя она и эффективна в некоторых областях, ей может быть трудно справиться с более сложными и тонкими задачами, требующими расширенных возможностей рассуждения.

Qwen2.5 7B набирает 35 баллов, что свидетельствует о значительном снижении производительности по сравнению с Qwen2.5 14B, подтверждая идею о том, что размер параметра играет критическую роль в сложных задачах на рассуждение, подобных тем, что представлены в BBH. Уменьшение возможностей этой модели, вероятно, ограничивает ее способность к обобщению и решению более абстрактных задач.

Модель Mistral NeMo 12B, набравшая 30 баллов, демонстрирует умеренно хорошие результаты в BBH, но испытывает трудности при решении более сложных задач на рассуждение по сравнению с моделями Phi-4 и Qwen аналогичного размера. Относительно низкие результаты говорят о том, что архитектура или процесс обучения Mistral NeMo 12B, возможно, не оптимизированы для решения высокоабстрактных задач рассуждения, вместо этого приоритет отдается более универсальному пониманию языка.

Llama 3 8B и Qwen2.5 3B набрали 27 и 26 баллов соответственно, показав, что обе модели испытывают трудности с задачами BBH, которые требуют более глубоких и сложных рассуждений. Результаты показывают, что эти модели, хотя и способны решать более простые языковые задачи, имеют ограничения, когда речь идет об абстрактном мышлении и умозаключениях более высокого порядка.

Llama 3.2 3B следует вплотную за ними с результатом 24 балла, демонстрируя те же недостатки, что и другие модели с меньшим количеством параметров. Ее относительно более низкий балл подчеркивает трудности, с которыми сталкиваются маленькие модели при решении задач, требующих обобщения в различных, абстрактных контекстах.

Gemma 2 2B, набравшая 18 баллов, демонстрирует трудности, с которыми сталкиваются очень маленькие модели при решении сложных рассуждений. Значительный разрыв в результатах между Gemma 2 2B и моделями с более высокими показателями говорит о том, что возможности этой модели недостаточны для решения задач, требующих более сложных навыков решения проблем.

Выводы

В данном исследовании изучалась производительность различных малопараметрических моделей большого языка в ряде бенчмарков, включая MMLU-PRO, GPQA, IFEval, MATH и BBH, с использованием GPT-4o в качестве опорной точки. Результаты наглядно демонстрируют ключевую роль размера и архитектуры модели в определении производительности в различных типах задач. Хотя небольшие модели, такие как Phi-4 14B, Qwen2.5 14B и Gemma 2 9B, демонстрируют высокие результаты в определенных областях, они постоянно отстают от более крупных моделей, таких как GPT-4o, особенно в более сложных рассуждениях и математических задачах.

В бенчмарках, требующих высокоуровневых рассуждений, таких как BBH, более мелкие модели, как правило, не смогли достичь уровня, сопоставимого с GPT-4o. Несмотря на то, что Phi-4 14B и Qwen2.5 14B демонстрируют потенциал для работы с абстрактными рассуждениями, они отстают, демонстрируя, что более высокая производительность модели часто коррелирует с улучшением обобщения задач и способности решать проблемы.

Общие выводы подчеркивают важность баланса между размером модели, архитектурой и оптимизацией обучения при разработке небольших SLM для специализированных случаев использования. Хотя компактные модели могут достичь заметных успехов в некоторых бенчмарках,

масштабирование параметров и улучшение архитектурного дизайна остаются крайне важными для задач, требующих глубоких рассуждений, абстракции и обобщения в нескольких областях.

Список литературы

1 Языковые модели Llama. [Электронный ресурс]. Доступно на: https://www.llama.com/

2 Языковые модели Gemma. [Электронный ресурс]. Доступно на: https://blog.google/technology/developers/google-gemma-2/

3 Языковые модели Qwen. [Электронный ресурс]. Доступно на: https://qwen2.org/qwen2-5/

4 Языковые модели Phi. [Электронный ресурс]. Доступно на: https://techcommunity.microsoft.com/blog/aiplatformblog/introducing-phi-4-microsoft%E2%80%99s-newest-small-language-model-specializing-in-comple/4357090

5 Языковые модели Mistral. [Электронный ресурс]. Доступно на: https://mistral.ai/en/news/mistral-nemo

6 Hendrycks D., Burns C., Basart S., Zou A., Mazeika M., Song D., Steinhardt J., Measuring massive multitask language understanding (MMLU), International Conference on Learning Representations, 2021, [Электронный ресурс]. Доступно на: https://openreview.net/pdf?id=d7KBjmI3GmQ

7 Rein D., Hou B.L., Stickland A.C., Petty J., Pang R.Y., Dirani J., Michael J., Bowman S.R., GPQA: A Graduate-Level Google-Proof Q&A Benchmark, 2023, New York. [Электронный ресурс]. Доступно на: https://arxiv.org/pdf/2311.12022

8 Zhou J., Lu T., Mishra S., Brahma S., Basu S., Luan Y., Zhou D., Hou L., Instruction-Following Evaluation for Large Language Models, 2023. [Электронный ресурс]. Доступно на: https://arxiv.org/pdf/2311.07911

9 Hendrycks D., Burns C., Kadavath S., Arora A., Basart S., Tang E., Song D., Steinhardt J., Measuring Mathematical Problem Solving with the MATH Dataset, 2021. [Электронный ресурс]. Доступно на: https://arxiv.org/pdf/2103.03874

10 Suzgun M., Scales N., Schärli N., Gehrmann S., Tay Y., Chung H.W., Chowdhery A., Le Q.L., Chi E.H., Zhou D., Wei J., Challenging BIG-Bench tasks and whether chain-of-thought can solve them, 2022. [Электронный ресурс]. Доступно на: https://arxiv.org/pdf/2210.09261

И.Р. Дашкин, Г.Д. Когай

Шағын тілдік модельдер: үлкен тілдік модельдер дәуіріндегі тиімділік пен өнімділік айырбастауы

Бұл құжат ресурсты көп қажет ететін үлкен тіл үлгілеріне (LLM) өміршең балама болып табылатын 15 миллиардтан аз параметрлері бар шағын тіл үлгілерінің (SLM) тиімділігі мен өнімділігін зерттейді. Llama 3.2 3B және Llama 3 8B, Gemma 2B және 9B, Qwen2.5 3B, 7B және 14B, Phi-4 14B және Mistral NeMo 12B сияқты заманауи SLM-лер стандартталған эталондар (MMLU-PRO, GPQA, IFEeva мәтіні, олардың жауаптары, жауаптары) арқылы салыстырылады логикалық ойлау және ойлау. Нәтижелер кейбір SLMs GPT-4o сияқты жоғары параметрленген үлгілерге жақын өнімділікті айтарлықтай төмен есептеу құнымен көрсететінін көрсетеді. Жұмыс зерттеушілер мен әзірлеушілерге практикалық ұсыныстар ұсына отырып, қол жетімді және экологиялық таза AI шешімдерін жасау үшін SLM әлеуетін көрсетеді.

Түйін сөздер: Шағын тіл үлгілері, шағын параметр үлгілері, табиғи тілді өңдеу, тіл үлгісін салыстыру, Llama 3.2 3B және Llama 3 8B, Gemma 2B және 9B, Qwen2.5 3B 7B және 14B, Phi-4 14B, Mistral NeMo 12B, энергия тиімділігі, өнімділік.

Раздел 3. «IT-технологии, энергетика, автоматизация и вычислительная техника»

I.R. Dashkin, G.D. Kogay

Small Language Models: The Efficiency-Performance Tradeoff in the Era of Large Language Models

The article explores the trade-off between efficiency and performance of small language models (SLMs) with fewer than 15 billion parameters, which represent a relevant alternative to resource-intensive large language models (LLMs). A comparison of modern SLMs, such as Llama 3.2 3B and Llama 3 8B, Gemma 2B and 9B, Qwen2.5 3B, 7B, and 14B, Phi-4 14B, and Mistral NeMo 12B, is conducted using standardized benchmarks (MMLU-PRO, GPQA, IFEval, MATH, BBH) to evaluate their capabilities in text generation, summarization, question answering, and logical reasoning. The results show that some SLMs demonstrate performance close to that of high-parameter models like GPT-4o, with significantly lower computational costs. The work highlights the potential of SLMs for creating more accessible and environmentally friendly solutions in the field of artificial intelligence, offering practical recommendations for researchers and developers.

Keywords: Small language models, low-parameter models, natural language processing, comparison of language models, Llama 3.2 3B and Llama 3 8B, Gemma 2B and 9B, Qwen2.5 3B 7B and 14B, Phi-4 14B, Mistral NeMo 12B, energy efficiency, performance.

References

1 Llama Language Models. [Electronic resource]. Available at: https://www.llama.com/

2 Gemma Language Models. [Electronic resource]. Available at: https://blog.google/technology/developers/google-gemma-2/

3 Qwen Language Models. [Electronic resource]. Available at: https://qwen2.org/qwen2-5/

4 Phi Language Models. [Electronic resource]. Available at: https://techcommunity.microsoft.com/blog/aiplatformblog/introducing-phi-4-microsoft%E2%80%99s-newest-small-language-model-specializing-in-comple/4357090

5 Mistral Language Models. [Electronic resource]. Available at: https://mistral.ai/en/news/mistral-nemo

6 Hendrycks D., Burns C., Basart S., Zou A., Mazeika M., Song D., Steinhardt J., Measuring massive multitask language understanding (MMLU), International Conference on Learning Representations, 2021, [Electronic resource]. Available at: https://openreview.net/pdf?id=d7KBjmI3GmQ

7 Rein D., Hou B.L., Stickland A.C., Petty J., Pang R.Y., Dirani J., Michael J., Bowman S.R., GPQA: A Graduate-Level Google-Proof Q&A Benchmark, 2023, New York. [Electronic resource]. Available at: https://arxiv.org/pdf/2311.12022

8 Zhou J., Lu T., Mishra S., Brahma S., Basu S., Luan Y., Zhou D., Hou L., Instruction-Following Evaluation for Large Language Models, 2023. [Electronic resource]. Available at: https://arxiv.org/pdf/2311.07911

9 Hendrycks D., Burns C., Kadavath S., Arora A., Basart S., Tang E., Song D., Steinhardt J., Measuring Mathematical Problem Solving with the MATH Dataset, 2021. [Electronic resource]. Available at: https://arxiv.org/pdf/2103.03874

10 Suzgun M., Scales N., Schärli N., Gehrmann S., Tay Y., Chung H.W., Chowdhery A., Le Q.L., Chi E.H., Zhou D., Wei J., Challenging BIG-Bench tasks and whether chain-of-thought can solve them, 2022. [Electronic resource]. Available at: https://arxiv.org/pdf/2210.09261