

Проект Galaxy от SenseTime нацелен на масштабирование отечественных ИИ-чипов

Источник: AI News Healthcare

Автор: Yang Fan

Дата публикации: 2026-07

Оригинал: <https://www.artificialintelligence-news.com/news/sensetimes-galaxy-project-targets-domestic-ai-chip-scale-up/>

ИИ-инфраструктура

вычислительные мощности

масштабирование

полупроводники

Компания SenseTime запустила проект **Galaxy Project**, объединив усилия с почти 20 партнерами для масштабирования отечественной инфраструктуры ИИ-чипов в Китае.

В программном докладе под названием «Интеллектуальная трансформация и симбиоз» Ян Фань — соучредитель компании и президент группы бизнеса крупногабаритных устройств — представил то, что SenseTime описывает как замкнутый цикл, соединяющий технологии на уровне чипов, партнерства в рамках экосистемы и коммерческое развертывание для отечественных вычислительных мощностей ИИ.

Наряду с проектом **Galaxy Project**, SenseTime подписала соглашение о космических вычислениях с производителем спутников Guoxing Aerospace и заключила исследовательское партнерство с пятью институтами, включая Шанхайскую лабораторию искусственного интеллекта, направленное на применение научных вычислений.

Ян обосновал выбор времени тремя сходящимися тенденциями: ростом спроса на токены в корпоративном секторе, догоняющим развитием промышленного ИИ по сравнению с потребительскими сценариями использования и достижением этапа коммерциализации отечественных чипов, когда центры интеллектуальных вычислений на базе китайских полупроводников могут развертываться быстрыми темпами.

Однако то, будет ли это «окно возможностей» таким открытым, как утверждает SenseTime, во многом зависит от показателей, которые компания не подвергала независимой проверке.

Показатели пропускной способности токенов сопровождаются большими оговорками

SenseTime заявляет, что её платформа крупномасштабных устройств в настоящее время обрабатывает в среднем 2,42 триллиона токенов ежедневно, и компания прогнозирует, что этот показатель вырастет в 25 раз до 10 триллионов токенов в день к четвертому кварталу 2026 года. Это прогноз, а не измеренный результат, и корпоративным покупателям, оценивающим инфраструктуру SenseTime, следует относиться к нему именно так, пока не начнут поступать квартальные отчеты.

Заявления об экономической эффективности, связанные с этим ростом, также являются самоотчетами. SenseTime утверждает, что её технология гетерогенного гибридного вывода обеспечивает увеличение показателя **Model FLOPs Utilisation** (коэффициент использования операций с плавающей запятой модели) на 85–152% на основных отечественных чипах, наряду с экономической эффективностью вывода, которую компания оценивает в 1,25 раза выше, чем у компонентов Nvidia серии H.

По сравнению с отечественными гомогенными конфигурациями вывода, SenseTime заявляет о 25-кратном увеличении выхода токенов при эквивалентных затратах — скачок, который, по словам компании, выводит оптимизированные гибридные кластеры вывода за пределы того, что индустрия ранее считала минимальным порогом рентабельности для отечественных вычислительных мощностей.

Ни один из этих показателей не сопровождается сторонним бенчмаркингом (тестированием производительности), а разрыв между оптимизированным тестовым кластером вендора и производственной средой заказчика — с её неравномерными конвейерами данных и задержками обновлений прошивки — обычно является тем местом, где такие цифры снижаются.

Заявления об адаптивности и проблема мультичиповых систем

Отечественные ИИ-чипы исторически сталкивались с проблемой фрагментированного программного стека: модели, обученные для одной архитектуры, часто требуют переработки для работы на другой. SenseTime заявляет, что создала полностековый уровень адаптации, охватывающий модели, фреймворки, операторы, инструментальные цепочки и аппаратное обеспечение, чтобы решить эту проблему и позволить клиентам переносить рабочие нагрузки между отечественными поставщиками чипов без масштабного переписывания кода.

Компания приводит два прикладных примера. В рабочей нагрузке **AI4S** (ИИ для науки) по предсказанию длинных белковых последовательностей SenseTime утверждает, что оптимизация объединенных операторов сократила общее время предсказания в три раза. В генерации видео с помощью **AIGC** (ИИ-генерируемый контент) компания заявляет о 93-процентном коэффициенте многокарточного параллельного ускорения для отечественных чипов, работающих с моделями **DiT** (Diffusion Transformer), наряду с тем, что она называет «бесплатной миграцией» для основных инструментов разработки ИИ.

Это те типы показателей, которые хорошо выглядят в тестовой «песочнице», но приобретают гораздо большее значение, когда они проходят стресс-тестирование на реальных клиентских конвейерах, работающих на смешанном оборудовании разных поколений.

Энергетические метрики получают новое название стандарта

SenseTime представила метрику, которую она называет **Tokens Per Watt** (Токенов на Ватт), позиционируемую как новый эталон для измерения эффективности центров обработки данных ИИ, наряду с **Computing-Power Collaboration Agent** (Агентом сотрудничества вычислительных мощностей), который управляет планированием ресурсов, прогнозированием цен на электроэнергию и оптимизацией хранения энергии в рамках того, что компания описывает как восьмьюровневую систему данных с пятью цепочками принятия решений.

Комбинируя вычисления, цены на электроэнергию и автоматизированное планирование, SenseTime заявляет об 80-процентном увеличении выхода токенов на единицу затрат электроэнергии, средних ценах на электроэнергию на 10% ниже, чем в сопоставимых региональных дата-центрах, и точности прогнозирования вычислительной нагрузки на уровне 96%.

Эти заявления заслуживают наблюдения в течение следующих нескольких кварталов, а не принятия на веру. Арбитраж цен на электроэнергию и точность прогнозирования нагрузки, как правило, ведут себя иначе, когда система проходит полный сезонный цикл с реальной волатильностью спроса, а не в условиях, в которых вендор обычно проводит свои пилотные испытания.

Впечатляющий список партнеров: от производителей чипов до поставщиков компонентов

Заявленная экосистема проекта **Galaxy Project** включает отечественных поставщиков чипов **Cambricon, Muxi, Hygon, Huawei Ascend, Moore Threads, Sunrise** и **Biren Technology**, партнера по компонентам **Xizhi Technology**, а также инфраструктурные фирмы, включая **Silicon Motion, Qujing Technology, Zhongke Jiahe, Qingcheng Jizhi, Sophon Information** и **Jiliu Technology**.

SenseTime заявляет, что план охватывает строительство одной «фабрики токенов», пяти вычислительных кластеров масштаба «10 000 калорий», совместную работу по десяти технологическим направлениям и поддержку 200 ИИ-стартапов.

«Отечественное производство — это не просто замена отдельных чипов, а скорее коллективные усилия по всей цепочке инновационных возможностей Китая: от чипов и компонентов до инфраструктуры и сценариев применения», — сказал Ян.

Ставки на космические, оптические и квантовые вычисления ориентированы на долгосрочную перспективу

Помимо краткосрочной инфраструктуры, SenseTime наметила работу над оптическими вычислениями для повышения эффективности дата-центров, применением квантовых вычислений в оптимизации ИИ и партнерством в

области космических вычислений с Guoxing Aerospace для создания того, что две компании называют **SenseTime Space Computing Constellation** (Космическая вычислительная созвездие SenseTime).

План SenseTime предусматривает первый запуск спутника в 2026 году с целью создания к 2030 году тысяч вычислительных спутников и вычислительной мощности в десятки тысяч петабайт.

Ян аргументировал, что ценность выходит за рамки чистых мощностей, представляя космические вычисления как способ расширить охват китайских ИИ-сервисов в средах со слабым сетевым соединением, таких как морские операции и реагирование на стихийные бедствия, и, как следствие, поддержать международный экспорт китайского ИИ.

Эта цель на 2030 год находится в пятилетней перспективе, а развертывание спутниковых вычислений такого масштаба не имеет прецедентов, с которыми можно было бы сопоставить сроки.

Физическая инфраструктура: от Шанхая до Эр-Рияда

На земле SenseTime заявляет, что её объект в Шанхае является первым в стране дата-центром, классифицированным по уровню интеллектуальных вычислений «5A», обрабатывающим более 20 триллионов токенов ежедневно в более чем 20 отраслях. Объект в Яньчэне был запущен с начальной мощностью 3000 петафлопс, ориентированной на применение в энергетике, производстве и экономике малых высот.

В Гонконге SenseTime строит то, что она описывает как крупнейший в регионе отечественный центр интеллектуальных вычислений, стремясь достичь 40 000 петафлопс к 2030 году. Компания также планирует создание того, что она называет первым зарубежным отечественным вычислительным кластером в Саудовской Аравии, позиционируемым как полностековая отечественная вычислительная база для Ближнего Востока.

В исследовательской области сотрудничество SenseTime с Шанхайской лабораторией ИИ, Пекинской академией Чжунгуаньцунь, Шэньчжэньской академией Хэтао, Шанхайским научно-исследовательским институтом инноваций в области алгоритмов и школой ИИ Шанхайского университета Цзяо Тун направлено на создание общей платформы, охватывающей вычисления, инструментарий и возможности моделей для исследований в области наук о жизни, материаловедения и производства. Ян назвал **AI for Science** (ИИ для науки) «ключевым рычагом парадигмальных инноваций в

фундаментальных исследованиях», связывая эту инициативу с более широкой государственной политикой Китая «Artificial Intelligence+» (Искусственный интеллект плюс).

Прогноз SenseTime о 10 триллионах токенов в день к четвертому кварталу 2026 года — это показатель, за которым следует следить и сравнивать его с тем, что компания сообщит по факту закрытия этого квартала.

Перевод выполнен: 22.07.2026 | ai4med.ru

Машинный перевод. Рекомендуем сверять с оригиналом при клиническом использовании.