

Обобщающая способность больших языковых моделей

Правда ли LLM обобщаются? Метрики, проблемы и границы

Начать презентацию →

Центральный вопрос

Действительно ли LLM обобщаются?

Или это просто очень сложное запоминание?



Этот вопрос определяет будущее ИИ

Признаки запоминания

● Что указывает на запоминание?

Дословное воспроизведение

- Точные цитаты из Шекспира, Википедии
- Утечка персональных данных из обучающей выборки

Заполнение шаблонов

- Стандартные ответы по образцу
- Отсутствие понимания причинности

Поиск поверхностных паттернов

- Работает только на знакомых форматах
- Ломается при малых изменениях входа



Выглядит как интеллект, но ломается на новых данных

Признаки настоящего обобщения

● Что указывает на обобщение?

Композиционное рассуждение

- "Переведи эту поэму на клингонский в формате хайку"
- Комбинация известных концептов новыми способами

Причинное понимание

- "Если гравитация была бы в два раза сильнее..."
- Логические цепочки от посылки к следствию

Адаптация к новым задачам

- Решение неизвестных задач без примеров
- Перенос знаний между областями



Решает задачи, которых модель никогда не видела

Классический пример: задача со словом "strawberry"

Вопрос к модели: "Сколько букв 'r' в слове strawberry?"

Ответы моделей:

- **Claude 3 Sonnet:** 2 буквы (неправильно)
- **GPT 4.5:** 2 буквы (неправильно)
- **Более старые модели:** 2-4 буквы

Правильный ответ: 3 буквы 'r' strawberry

Что показывает пример:

Проблема токенизации

- Модель видит токены, не буквы
- "strawberry" = один токен
- Сложно анализировать внутреннюю структуру

Границы обобщения

- Простая задача для человека
- Сложная для современных LLM

Новые формы переобучения

Переобучение на тесты

- Модели "заучивают" популярные бенчмарки
- 85%+ на тесте
- Плохо на похожих задачах
- Различие обучение/тест исчезает

Временное переобучение

- Знания "заморожены" на дату среза
- Выдумки о событиях после 2022-2023
- Неспособность обновлять знания
- Устаревшая информация

Переобучение на распределение

- Работает на "интернет-стиле" текстов
- Плохо на формальных документах
- Культурные предрассудки
- Англоязычный bias

Эмерджентные способности: наблюдения

Что мы наблюдаем?

Арифметика

- Маленькие модели: не могут сложить двузначные числа
- Большие модели: решают сложную арифметику
- Резкий переход в районе 10-50 миллиардов параметров

Пошаговое рассуждение

- Маленькие модели: не могут рассуждать пошагово
- Большие модели: естественно делают это
- Появляется "внезапно" при масштабировании

Обучение в контексте

- Способность учиться из примеров в запросе
- Эффект порога: либо есть, либо нет

Эмерджентные способности: критика

"Мираж эмерджентных способностей" (Schaeffer et al., 2023)

Гипотеза: внезапные способности — артефакт метрик

- Нелинейные метрики создают иллюзию скачка
- На самом деле: плавное улучшение
- Пример: точность < 80% считается провалом

Настоящие внезапные способности?

- Пошаговое рассуждение: порог действительно резкий
- Обучение в контексте: качественное изменение поведения
- Композиционность: новые комбинации способностей

Композиционное обобщение

Что это такое?

Способность комбинировать известные элементы новыми способами

Примеры тестов:

- Обучение: "прыгни дважды"
- Тест: "прыгни трижды"
- Может ли модель обобщить?



LLM плохо справляются с систематической композиционностью

Загадка нейросетей: почему одно получается, а другое — нет?

Исследование на простых картинках

Проблема: Stable Diffusion создаёт фантастических существ, но не может нарисовать синюю панду

Эксперимент:

- Учили нейросеть рисовать простые фигуры: ▲ ● , ○ ● , ■ ●
- Показывали только некоторые комбинации при обучении
- Проверяли: сможет ли нарисовать новые комбинации?

Пример: Учили на красных треугольниках и синих кругах. Сможет ли нарисовать синий треугольник?

Почему способности появляются внезапно?

Открытие: способности перемножаются, а не складываются!

Простая аналогия:

Представь, что нейросеть учится готовить. Чтобы сделать **синий пирог**, нужно уметь:

- Готовить пирог (70% успеха)
- Делать синий краситель (80% успеха)

Итоговый успех: $70\% \times 80\% = 56\%$ (плохо!)

Но если каждый навык улучшится до 90%: **Новый успех:** $90\% \times 90\% = 81\%$ (хорошо!)

Поэтому способности появляются скачками:

- Сначала ничего не работает (56%)

Редкое учится дольше

Конкретный пример:

В обучающих данных было:

- **1000 фото** белых людей в костюмах
- **10 фото** темнокожих людей в костюмах

Результат:

- Нейросеть хорошо рисует белых людей в костюмах
- Плохо рисует темнокожих людей в костюмах
- Очень плохо рисует темнокожих людей в традиционной одежде

Почему так происходит?

Популярные комбинации

Учатся быстро

Редкие комбинации

Нужно учить в 10 раз дольше!



⚠ Если остановить обучение рано → дискриминация редких групп

Устойчивость к возмущениям

Типы тестов устойчивости:

Adversarial устойчивость

- Малые возмущения входа
- Замена синонимами
- Перефразирование

Калибровка уверенности

- Коррелирует ли уверенность с точностью?
- Ошибка ожидаемой калибровки



⚠️ LLM слишком уверены в себе и хрупкие к изменениям

Перенос знаний между областями

Типы переноса:

Перенос без примеров

- Между языками: английский → китайский
- Между областями: медицина → право

Адаптация с несколькими примерами

- Сколько примеров нужно для новой задачи?
- Эффективность = точность / количество примеров



Перенос знаний работает лучше всего между похожими задачами

Проблема насыщения тестов

Старые тесты достигли потолка

Примеры "решенных" тестов:

- GLUE (2018): достигнут человеческий уровень
- SuperGLUE (2019): почти насыщен
- MMLU (2020): Claude 4 показывает 86%

Последствия насыщения:

- Невозможно различить топовые модели
- Подгонка под формат конкретного теста
- Ложное чувство прогресса в области

Почему тесты ломаются?

Закон Гудхарта в действии:

"Когда метрика становится целью, она перестаёт быть хорошей метрикой"

Основные проблемы:

Загрязнение данных

- Тестовые данные попадают в обучающую выборку
- Модель "знает" правильные ответы

Переобучение на формат

- Модели учатся "взламывать" формат теста
- Множественный выбор против открытых вопросов

Ложные корреляции

- Модели находят shortcuts вместо настоящего решения

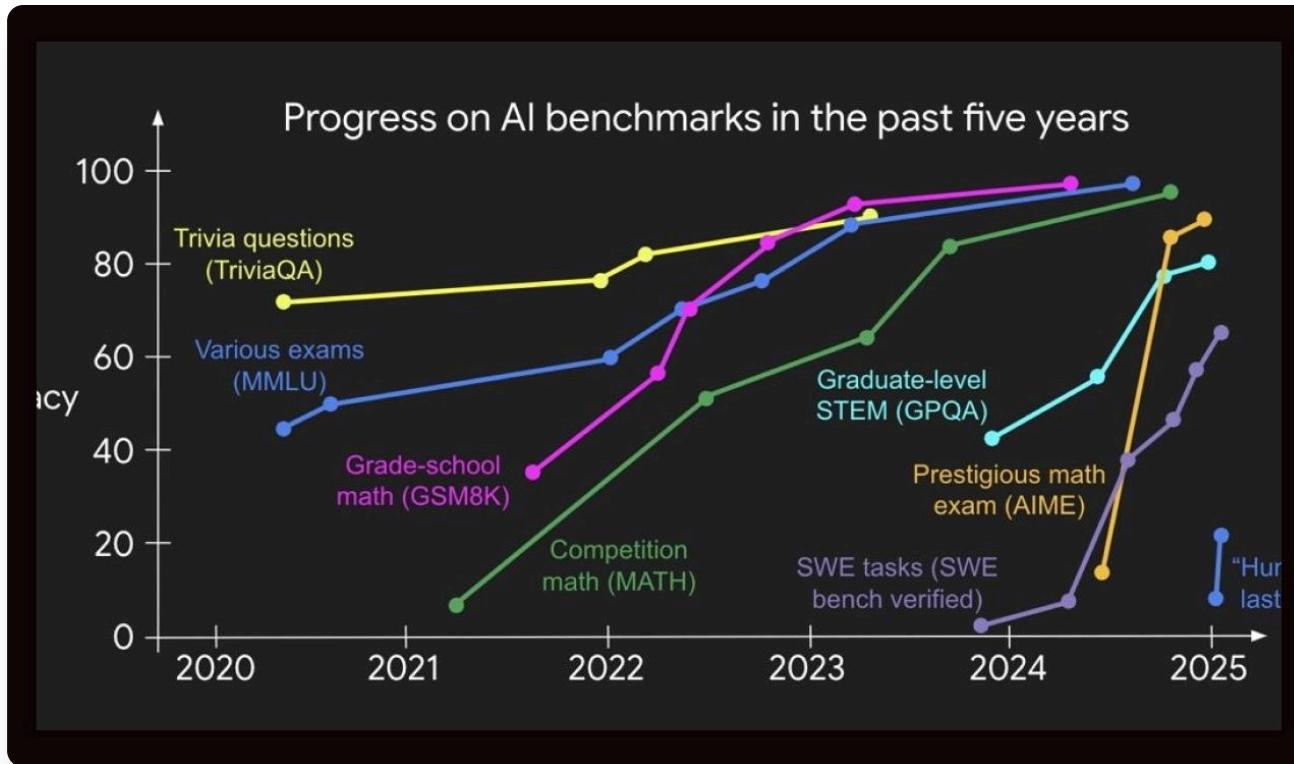
Математические тесты: AIME

AIME (Американские математические олимпиады)

Что это:

- 30 задач олимпиадного уровня
- Требует глубокого математического мышления
- Тестирует рассуждения, не запоминание

Рост бенчмарков во времени



Текущие бенчмарки

	Claude Opus 4	Claude Sonnet 4	Claude Sonnet 3.7	OpenAI o3	OpenAI GPT-4.1	Gemini 2.5 Pro <i>Preview (05-06)</i>
Agentic coding <i>SWE-bench</i> <i>Verified</i> ^{4,5}	72.5% / 79.4%	72.7% / 80.2%	62.3% / 70.3%	69.1%	54.6%	63.2%
Agentic terminal coding <i>Terminal-bench</i> ^{2,5}	43.2% / 50.0%	35.5% / 41.3%	35.2%	30.2%	30.3%	25.3%
Graduate-level reasoning <i>GPQA Diamond</i> ⁵	79.6% / 83.3%	75.4% / 83.8%	78.2%	83.3%	66.3%	83.0%
Agentic tool use <i>TAU-bench</i>	Retail 81.4%	Retail 80.5%	Retail 81.2%	Retail 70.4%	Retail 68.0%	—
	Airline 59.6%	Airline 60.0%	Airline 58.4%	Airline 52.0%	Airline 49.4%	—
Multilingual Q&A <i>MMMLU</i> ³	88.8%	86.5%	85.9%	88.8%	83.7%	—
Visual reasoning <i>MMMU</i> (validation)	76.5%	74.4%	75.0%	82.9%	74.8%	79.6%
High school math competition <i>AIME 2025</i> ^{4,5}	75.5% / 90.0%	70.5% / 85.0%	54.8%	88.9%	—	83.0%

Разрыв простой против сложной математики

Простая математика

Школьный уровень

- Базовые операции
- Стандартные алгоритмы

Сложная математика

Университетский/олимпиадный уровень

- Творческие решения
- Многошаговые доказательства



Чем сложнее рассуждение, тем больше разрыв между LLM и людьми

Улучшенные классические тесты

MMLU-Pro

- Усложненная версия MMLU
- 10 вариантов ответа вместо 4
- Более сложные рассуждения

GPQA (уровень аспирантуры)

- Вопросы уровня PhD по наукам
- Биология, химия, физика
- Экспертные знания required

Новые специализированные тесты

SimpleQA (OpenAI)

- 4,326 коротких фактических вопросов
- Фокус на обнаружение галлюцинаций
- **o1-preview:** ~42.7% правильных ответов
- **GPT-4o:** ~38.2% правильных ответов
- **Claude 3.5 Sonnet:** ~28.9% правильных ответов

SWE bench (реальное программирование)

- 2,294 реальных GitHub issues из Python репозиториев
- Требует понимания больших кодовых баз
- Координация изменений в нескольких файлах
- **Лучшие модели:** 20-45% решенных задач
- **Claude 3.5 Sonnet:** ~49% на SWE-bench Verified

Adversarial тестирование: найденные слабости

Систематические слабости Claude 4:

Галлюцинации: 10-15% на фактических вопросах

Математические рассуждения: деградация на сложных задачах

Длинный контекст: производительность падает с ростом

Многошаговое планирование: ошибки накапливаются

Уязвимости к инъекциям: 1% успешных попыток взлома

Типы Adversarial тестов

Возмущения входа

- На уровне символов
- Семантически сохраняющие перефразы
- Межъязыковые вариации
- Замены синонимами

Сценарии вне распределения

- Необычные контексты
- Границочные случаи
- Противоречивые инструкции
- Новые комбинации областей

Систематическое стресс-тестирование

- Возрастающая сложность
- Ограничения ресурсов
- Мультимодальные вызовы
- Временные ограничения



LLM гораздо более хрупкие, чем показывают стандартные тесты

Проблемы автоматических метрик

Основные проблемы:

BLEU для генерации текста:

- Плохо коррелирует с качеством
- Не учитывает семантику
- Предрассудки к поверхностному сходству

Точное совпадение для вопросов:

- "42" против "сорок два" против "примерно 42"
- Все семантически правильные, но разный балл

Perplexity:

- Не коррелирует с качеством работы
- Лучший perplexity ≠ лучшая модель

Проблемы человеческой оценки

Основные вызовы:

Субъективность согласия оценщиков:

- Фактичность: высокое согласие (85%)
- Полезность: среднее согласие (62%)
- Креативность: низкое согласие (34%)

Проблемы масштабирования:

- Экспертная оценка: \$10 за оценку
- Краудсорсинг: \$0.50 за оценку
- Полная оценка Claude 4: \$500,000

Новые решения для оценки

LLM как судья:

- LLM оценивает другие модели
- Автоматизация экспертной оценки
- Проблема: bias к собственному стилю

Конституционный ИИ:

- Самооценка через принципы
- Итеративное улучшение
- Прозрачные критерии

Человек в цикле:

- Гибридные pipeline
- Комбинация автоматических и человеческих оценок
- Активное обучение для сложных случаев

Старые законы масштабирования

Законы Kaplan (2020):

Основная идея:

- Увеличивай параметры настолько, насколько хватает вычислений
- Данные вторичны
- Больше параметров = лучше результат

Пример GPT-3:

- 175 миллиардов параметров
- 300 миллиардов токенов
- Огромная модель для того времени

Результат: модель была недообучена!

Революция Chinchilla (2022)

Новое открытие:

Оптимальное соотношение: 20 токенов на параметр

Пример Chinchilla:

- 70 миллиардов параметров (меньше GPT-3)
- 1.4 триллиона токенов (больше GPT-3)
- Лучшая производительность при том же бюджете

Для GPT-3 нужно было 3.5 триллиона токенов (в 12 раз больше!)

Ограничения масштабирования

Вычислительная стена

- Экспоненциальный рост стоимости
- Ограниченная доступность чипов
- Энергетические ограничения
- Физические пределы

Стена данных

- Качественные данные кончаются
- Интернет уже собран
- Книги почти все оцифрованы
- Научные статьи исчерпаны

Убывающая отдача

- Diminishing returns от параметров
- Логарифмический рост качества
- Необходимость новых подходов
- Архитектурные инновации



⚠ Эра простого масштабирования подходит к концу

Традиционное масштабирование

Старый подход:

Больше при обучении:

- Больше параметров
- Больше данных для обучения
- Больше вычислений при обучении

Результат:

- Фиксированная производительность на выходе
- Нет контроля над качеством во время использования
- Одна попытка = один результат

Масштабирование во время теста

Новый подход:

Больше при использовании:

- Больше вычислений во время ответа
- Более длинные цепочки рассуждений
- Множественные попытки и самопроверка

Результат:

- Качество растет с бюджетом вычислений
- Контроль над качеством ответа
- Адаптивная точность под задачу

Методы масштабирования рассуждений

Цепочка рассуждений

- Пошаговое рассуждение
- "Давайте думать пошагово"
- Улучшение на 20-50%
- Прозрачный процесс

Самосогласованность

- Множественные пути рассуждения
- Голосование большинством
- Дополнительные 10-20% улучшения
- Устойчивость к ошибкам

Дерево мыслей

- Ветвящиеся стратегии
- Исследование подходов
- Откат при ошибках
- Систематический поиск

Механизм О поколения OpenAI

"Цепочка мыслей во время обучения"

Как это работает:

- Обучение с подкреплением на процесс рассуждения
- Награда за правильные промежуточные шаги
- Обучение "думать" перед ответом

Свойства масштабирования:

- Больше вычислений → лучшие результаты
- Более длинные рассуждения → выше точность
- Механизмы самокоррекции

Ограничения О поколения OpenAI

Стоимость и скорость:

- Значительно дороже обычных моделей
- Медленнее время ответа
- Ограниченная доступность

Надежность:

- Все еще ошибается на простых задачах
- Рассуждения не всегда понятны
- Может "зациклиться" на неверном пути

Масштабируемость:

- Работает не для всех типов задач
- Требует специальной настройки

Фундаментальные вопросы масштабирования рассуждений

Ключевые открытые вопросы:

Есть ли пределы?

- Убывающая отдача при длинных цепочках?
- Вычислительная сложность для реальных проблем
- Накопление ошибок в многошаговых рассуждениях

Обобщаемость паттернов?

- Работают ли стратегии рассуждения между областями?
- Перенос от математики к науке к здравому смыслу?
- Композиционные способности рассуждения

Практические вызовы масштабирования рассуждений

Соотношение затрат и выгод:

- В 10 раз больше вычислений за в 2 раза лучшую точность?
- Когда стоит использовать масштабирование рассуждений?
- Оптимальное распределение бюджета

Проблемы надежности:

- Более длинные цепочки → более хрупкие?
- Распространение галлюцинаций в рассуждении
- Калибровка уверенности для сложных рассуждений



Может ли масштабирование рассуждений привести к качественному прорыву?

Теория коллапса модели

Гипотеза о современных моделях:

Что происходит:

- Данные для обучения содержат значительную долю контента, созданного ИИ
- Модель обучена на выходах предыдущих поколений
- Результат: регрессия в творческих задачах

Механизм коллапса:

1. **Усиление ошибок:** ошибки накапливаются между поколениями
2. **Коллапс режимов:** снижение разнообразия выходов
3. **Сдвиг распределения:** текст ИИ отличается от человеческого

Экспериментальные доказательства коллапса

Shumailov et al. (2023):

- VAE на MNIST деградирует за 5-7 поколений
- Ухудшение качества неизбежно без инъекции человеческих данных

Признаки в современных моделях:

- Повторяющиеся фразы и паттерны
- "Безопасные" но неинспирированные ответы
- Потеря знаний с длинным хвостом
- Снижение метрик креативности

Контрмеры:

- Курируемые пайплайны данных с человеческим надзором
- Обнаружение ИИ для фильтрации синтетического контента

Состояние интернета в 2024

Состав интернет-контента:

- **Человеческий оригинал:** ~60% (и падает)
- **Сгенерированный ИИ:** ~25% (быстро растёт)
- **Совместный человеко-ИИ:** ~10%
- **Неизвестно:** ~5%

Проблемы качества:

- Выдуманные факты распространяются
- Шаблонный язык доминирует
- Культурная гомогенизация контента
- Потеря аутентичного голоса и нюансов

Исчерпание качественных данных

Оставшиеся качественные источники:

- **Книги:** ~5% не оцифрованы
- **Научные статьи:** ~90% уже использованы
- **Новости:** ~100% собраны
- **Код:** ~95% публичного кода использовано

Пути решения:

- **Мультимодальные данные:** видео, аудио, сенсорные
- **Симуляционные среды** для синтетических данных
- **Сотрудничество человека и ИИ** в создании контента
- **Качество превыше количества**



Небольшой курированный датасет лучше огромного шумного

Технические ограничения будущего

Мультимодальное обобщение

- Единое пространство представлений
- Межмодальные рассуждения
- Временная согласованность
- Пространственное понимание

Непрерывное обучение

- Преодоление катастрофического забывания
- Эластичные системы памяти
- Мета-обучение для адаптации
- Сохранение старых знаний

Вычислительные пределы

- Убывающая отдача от масштабирования
- Ограничения потребления энергии
- Аппаратные ограничения
- Физические пределы полупроводников

Фундаментальные вопросы

Понимание против имитации

- Как отличить настоящее понимание?
- Можно ли измерить "глубину" понимания?
- Проблема китайской комнаты для LLM
- Критерии настоящего интеллекта

Гарантии устойчивости

- Формальные гарантии безопасности
- Предсказуемое поведение в граничных случаях
- Калиброванная оценка неопределенности
- Робастность к adversarial атакам

Согласование с человеческими ценностями

- Обобщение морального рассуждения
- Культурная адаптивность
- Согласование ценностей при масштабировании
- Универсальные против относительных ценностей

Ключевой вопрос будущего

Достигнем ли мы искусственного общего интеллекта через масштабирование LLM?

Или нужны фундаментально новые подходы?

За масштабирование

- Эмерджентные способности при росте
- Масштабирование рассуждений показывает прогресс
- Путь к AGI может быть непрерывным

Против масштабирования

- Фундаментальные ограничения данных
- Проблемы с истинным пониманием
- Необходимость новых архитектур

Что мы знаем точно

Установленные факты

- Законы масштабирования работают
- Эмерджентные способности реальны (но могут быть постепенными)
- Качество данных важнее количества
- Тесты не отражают реальную производительность
- Масштабирование рассуждений дает новые возможности

Открытые вопросы

Что остается неясным ?

- Действительно ли LLM "понимают"?
- Как измерить истинное обобщение?
- Можно ли избежать коллапса модели?
- Есть ли пределы у подхода масштабирования?
- Приведет ли масштабирование рассуждений к AGI?



Будущее ИИ зависит от ответов на эти вопросы

Спасибо за внимание!

Вопросы и обсуждение