Лекция 7.

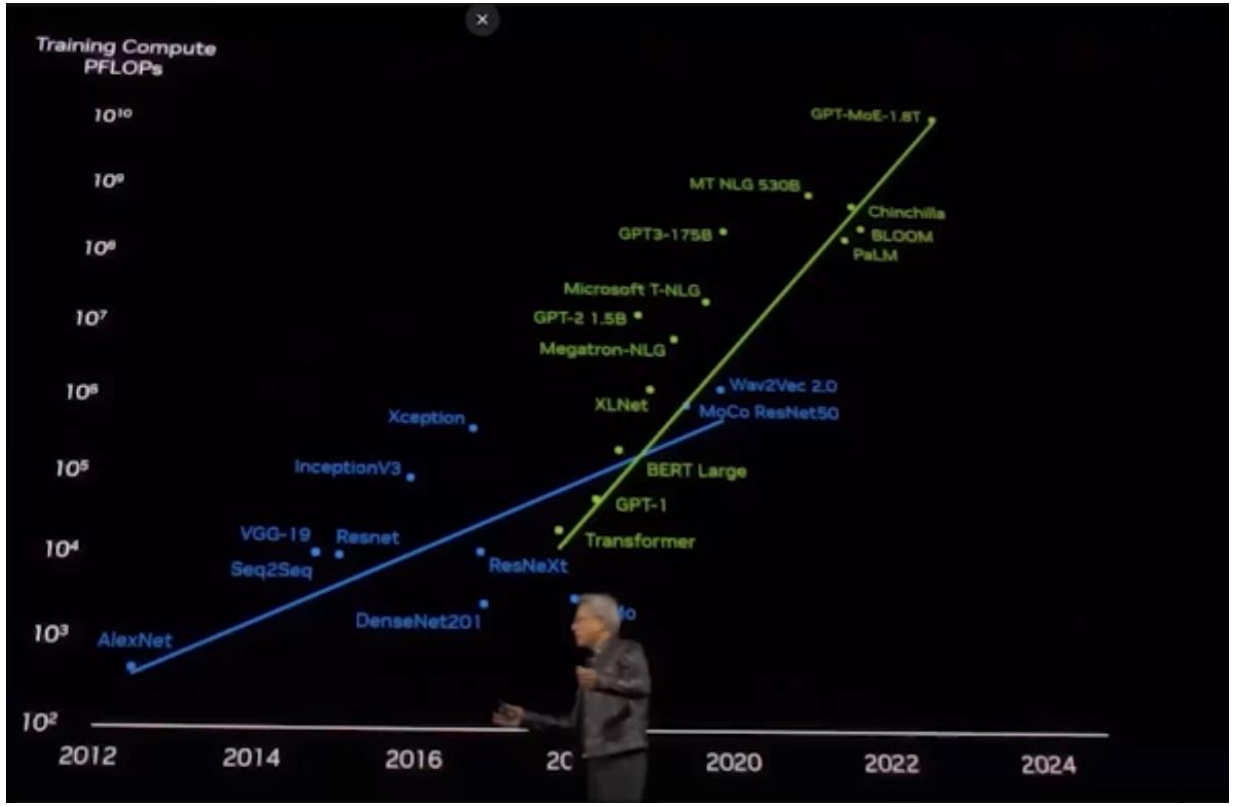
MoE, DeepSeek, Qwen3

MoE, DeepSeek, Qwen3

- MoE
- Mixtral 8x7B
- JetMoE
- DeepSeek
- Qwen3

Mixture of Experts (MoE)

- слухи про то, что GPT-4 – 1.8T MoE model с 16-ю экспертами, 111B каждый

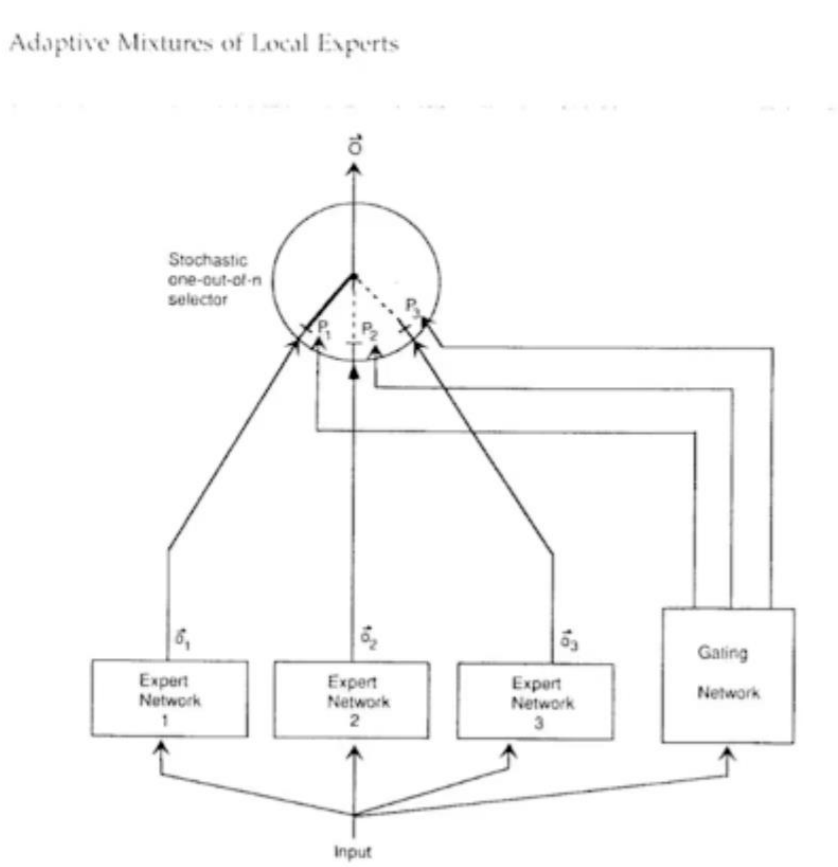


Mixture of Experts (MoE)

- Истоки подхода в 1991 году.

- Одинаковый вход идет в 3 слоя-эксперта и в Gating Network. **Gating Network (Сетевое устройство/сеть-маршрутизатор)** решает, какому «эксперту» (или каким экспертам) передать входные данные.

- Каждый эксперт должен отвечать за разные подмножества тренировочной выборки, специализируясь на них.



MoE: Sparsity

Sparsity (Разреженность) - это концепция в машинном обучении, где только небольшая часть параметров или активаций модели является активной в любой момент времени.

- **LSTM с экспертами** — это гибридная архитектура, сочетающая **LSTM (Long Short-Term Memory)** и **Mixture of Experts (MoE)**.

- Идея **Conditional computation.**

- Каждый «эксперт» -FeedForward, то есть представляет собой **полносвязную нейронную сеть (FeedForward Neural Network)** с одним или несколькими скрытыми слоями.

- Gating Network обучается и отвечает за выбор экспертов.

- Выбирается только top-k экспертов на каждый токен.

MoE: результаты

Table 2: Results on WMT'14 En→ Fr newstest2014 (bold values represent best results).

Model	Test Perplexity	Test BLEU	ops/timenstep	Total #Parameters	Training Time
MoE with 2048 Experts	2.69	40.35	85M	8.7B	3 days/64 k40s
MoE with 2048 Experts (longer training)	**2.63**	**40.56**	85M	8.7B	6 days/64 k40s
GNMT (Wu et al., 2016)	2.79	39.22	214M	278M	6 days/96 k80s
GNMT+RL (Wu et al., 2016)	2.96	39.92	214M	278M	6 days/96 k80s
PBMT (Durrani et al., 2014)		37.0			
LSTM (6-layer) (Luong et al., 2015b)		31.5			
LSTM (6-layer+PosUnk) (Luong et al., 2015b)		33.1			
DeepAtt (Zhou et al., 2016)		37.7			
DeepAtt+PosUnk (Zhou et al., 2016)		39.2			

Table 3: Results on WMT'14 En → De newstest2014 (bold values represent best results).

Model	Test Perplexity	Test BLEU	ops/timestep	Total #Parameters	Training Time
MoE with 2048 Experts	**4.64**	**26.03**	85M	8.7B	1 day/64 k40s
GNMT (Wu et al., 2016)	5.25	24.91	214M	278M	1 day/96 k80s
GNMT +RL (Wu et al., 2016)	8.08	24.66	214M	278M	1 day/96 k80s
PBMT (Durrani et al., 2014)		20.7			
DeepAtt (Zhou et al., 2016)		20.6			

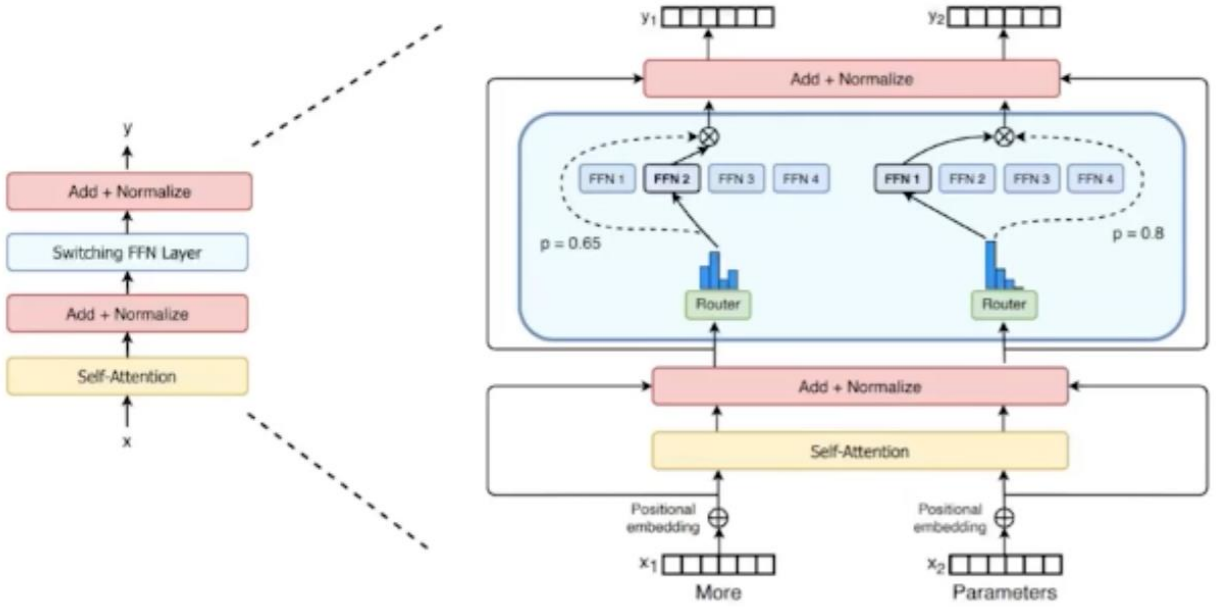
Table 4: Results on the Google Production En→ Fr dataset (bold values represent best results).

Model	Eval Perplexity	Eval BLEU	Test Perplexity	Test BLEU	ops/timestep	Total #Parameters	Training Time
MoE with 2048 Experts	**2.60**	**37.27**	**2.69**	**36.57**	85M	8.7B	1 day/64 k40s
GNMT (Wu et al., 2016)	2.78	35.80	2.87	35.56	214M	278M	6 days/96 k80s

Switch Transformer

Switch Transformer — это продвинутая версия Mixture of Experts от Google.

- Основан на архитектуре T5 (кодер-декодер)
- **Выбирается 1 эксперт**
- Вспомогательный loss для каждого Switch слоя



Switch Transformer

Преимущества Switch Transformer:

- **Эффективность**

В 7-10 раз быстрее обучения обычных плотных моделей

Меньше потребление памяти

Можно обучать модели с **триллионами параметров**

- **Качество**

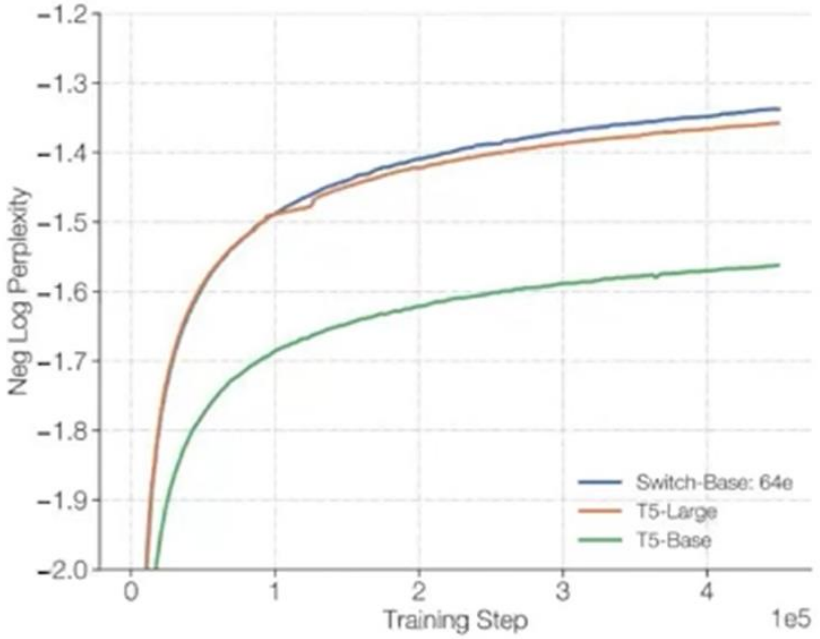
Сохраняет или улучшает качество относительно плотных моделей. Эксперты становятся более **специализированными**

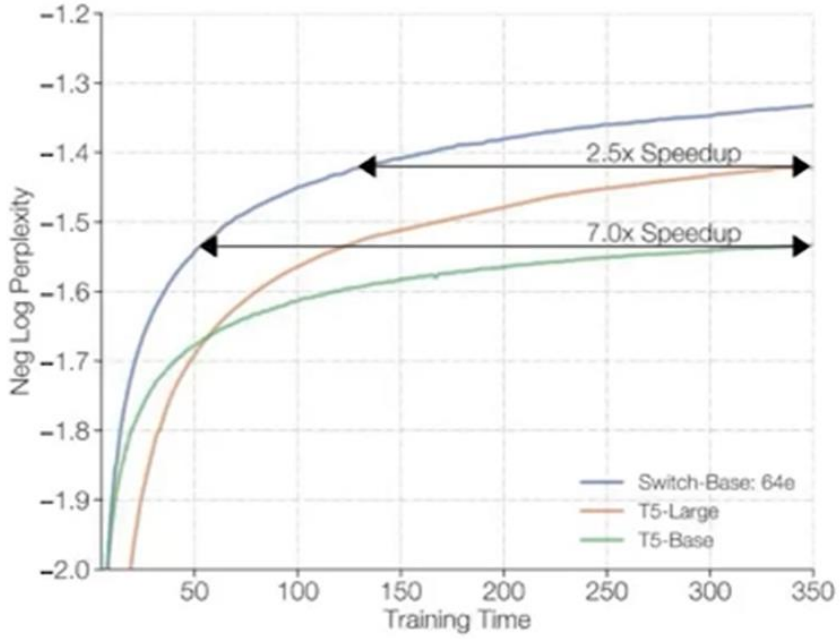
- **Балансировка нагрузки**

Умное распределение токенов между экспертами

Избегает "перегруженных" и "ленивых" экспертов

Switch Transformer: ускорение сходимости





Switch Transformer: что выбрать?

Model	Precision	Quality @100k Steps (↑)	Quality @16H (↑)	Speed (ex/sec) (↑)
Experts FF	float32	-1.548	-1.614	1480
Expert Attention	float32	-1.524	**-1.606**	1330
Expert Attention	bfloat16	[diverges]	[diverges]	–
Experts FF + Attention	float32	**-1.513**	-1.607	1240
Expert FF + Attention	bfloat16	[diverges]	[diverges]	–

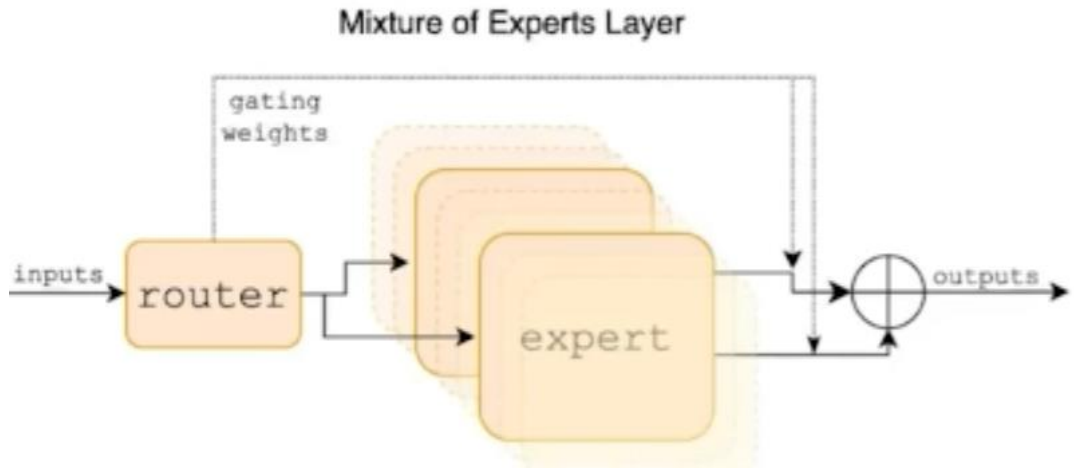
различные архитектурные варианты Switch Transformer в зависимости от того, какие компоненты сделаны экспертами

Что выучивают эксперты?

Expert specialization	Expert position	Routed tokens
Sentinel tokens	Layer 1	been <extra_id_4><extra_id_7>floral to <extra_id_10><extra_id_12><extra_id_15> <extra_id_17><extra_id_18><extra_id_19>...
	Layer 4	<extra_id_0><extra_id_1><extra_id_2> <extra_id_4><extra_id_6><extra_id_7> <extra_id_12><extra_id_13><extra_id_14>...
	Layer 6	<extra_id_0><extra_id_4><extra_id_5> <extra_id_6><extra_id_7><extra_id_14> <extra_id_16><extra_id_17><extra_id_18>...
Punctuation	Layer 2	, , , , , , , , , - , , , , ,).)
	Layer 6	, , , , , : , : , & , & & ? & - , , ? , , , , <extra_id_27>
Conjunctions and articles	Layer 3	The the the the the the the the the The the the the the the The the the the
	Layer 6	a and and and and and and and or and a and . the the if ? a designed does been is not
Verbs	Layer 1	died falling identified fell closed left posted lost felt left said read miss place struggling falling signed died falling designed based disagree submitted develop
Visual descriptions *color, spatial position*	Layer 0	her over her know dark upper dark outer center upper blue inner yellow raw mama bright bright over open your dark blue
Proper names	Layer 1	A Mart Gr Mart Kent Med Cor Tri Ca Mart R Mart Lorraine Colin Ken Sam Ken Gr Angel A Dou Now Ga GT Q Ga C Ko C Ko Ga G
Counting and numbers *written and numerical forms*	Layer 1	after 37 19. 6. 27 I I Seven 25 4, 54 I two dead we Some 2012 who we few lower each

Mixtral 8x7B

- 8 экспертов по 7b (в сумме).
- 2 активных за раз (К=2).
- Скорость как у 14b моделей.
- Длина последовательности: 32k.
- Мультиязычная.



Mixtral 8x7B. Архитектура Mixture of Experts

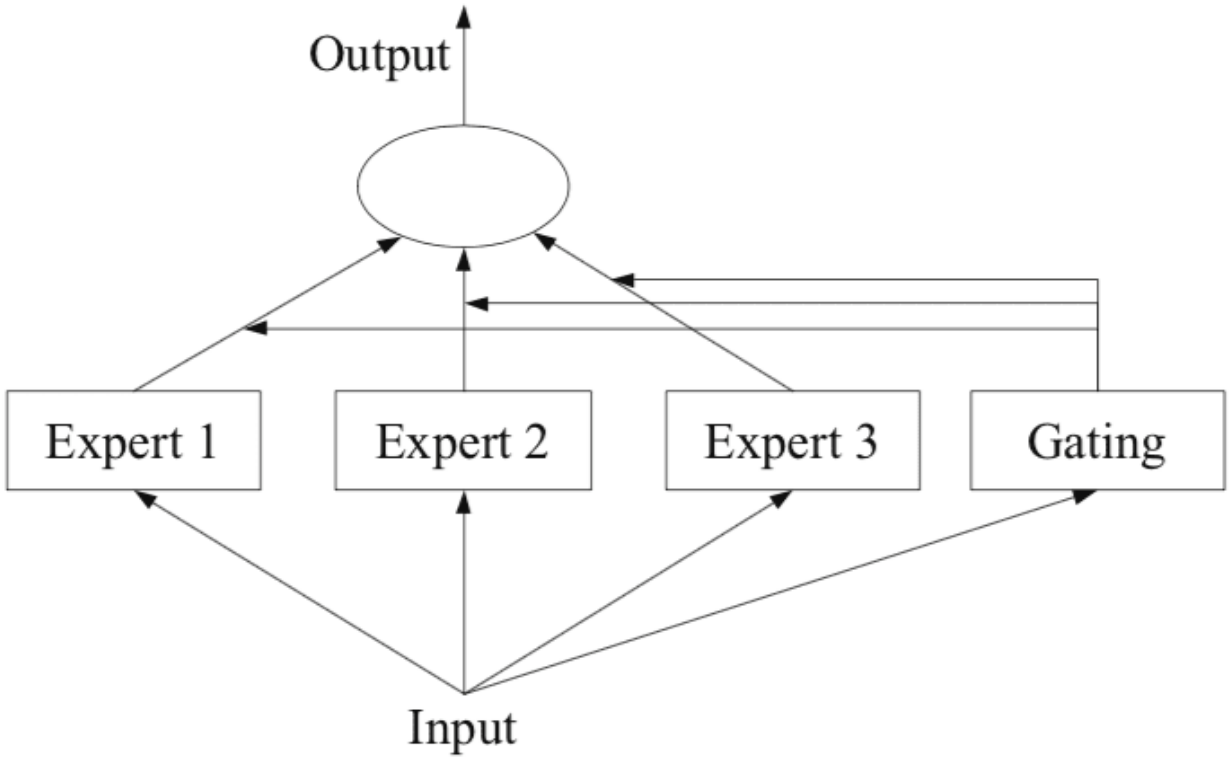
Некое множество “экспертов”, иными словами специализированных сетей, применяются для того, чтоб решать комплексную проблему по частям.
То есть есть несколько разных сетей, каждая из которых хорошо умеет решать свою узкую задачу, и когда в модель приходит запрос, то на этот запрос отвечают не все эксперты, а только некоторое их множество, а то и один.

Архитектура Mixture of Experts

Основные этапы создания mixture of experts модели:

- Поделить входящую в сеть задачу на подзадачи. При этом подзадачи могут пересекаться.
- Обучить экспертную сеть для решения каждой подзадачи.
- Использовать gating (routing) модель, что решить, какого эксперта мы используем. Gating model принимает поданный в модель контекст, на основе него оценивает все экспертные предсказания и выбирает, какой эксперт / эксперты будет давать итоговый ответ пользователю.
- Объединение предсказаний экспертов. Можно выбрать одного эксперта, который даст ответ, а можно выбрать нескольких и совместить их ответы.

Архитектура Mixture of Experts



Sparse Mixture of Experts

Но ведь мы можем сразу отсекать "неподходящих" экспертов. Здесь модель выбирает экспертов, к которым стоит обратиться, исходя из полученного контекста. Остальные эксперты не получат входных данных и не будут работать, время ответа сократится.

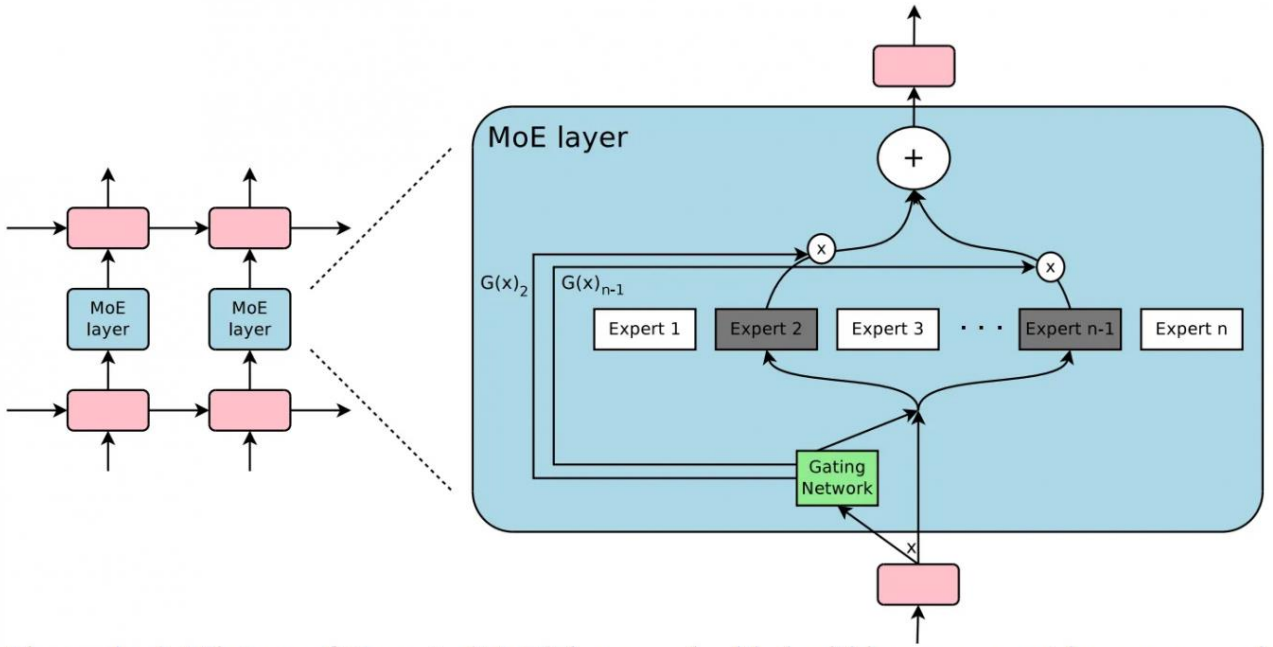
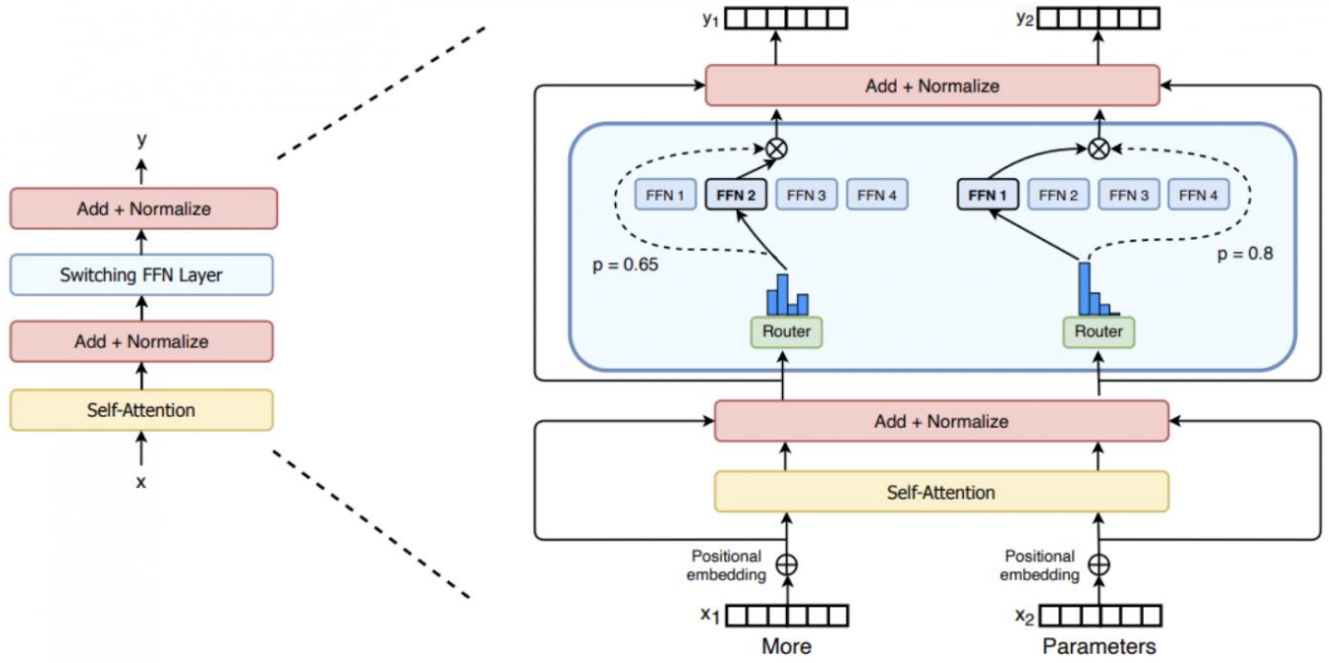


Figure 1: A Mixture of Experts (MoE) layer embedded within a recurrent language model. In this case, the sparse gating function selects two experts to perform computations. Their outputs are modulated by the outputs of the gating network.

Sparse Mixture of Experts

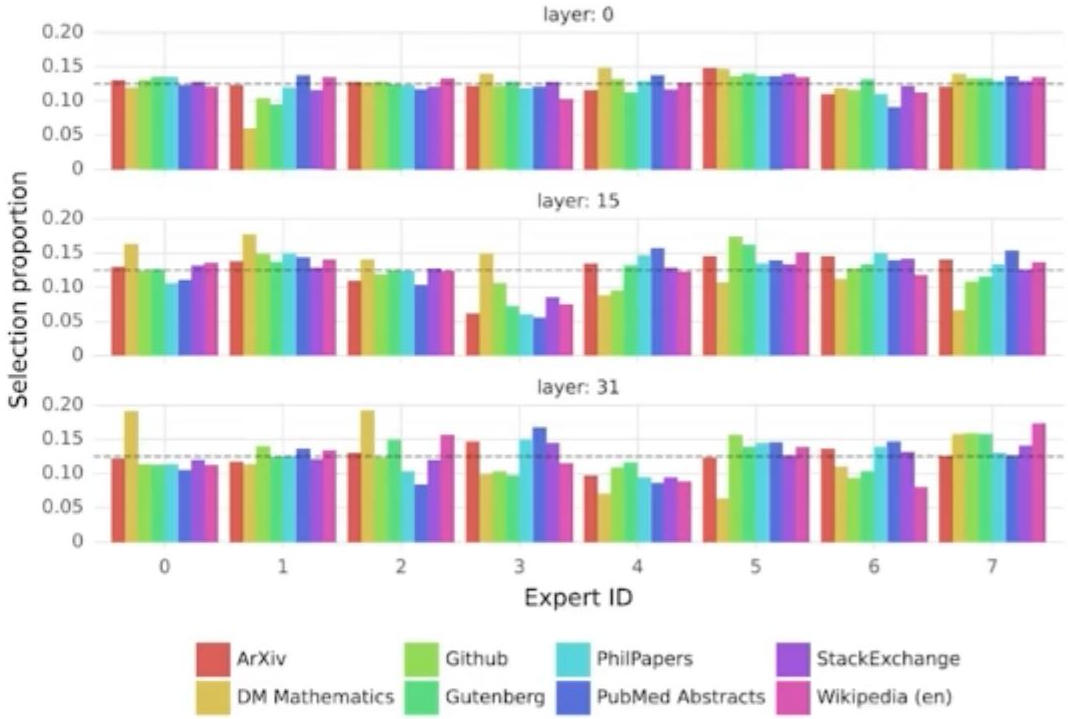


Разреженность знаний по набору экспертов позволяет запускать только части сложной системы, вычисления заметно ускоряются.

Mixtral 8x7B

	LLaMA 2 70B	GPT-3.5	Mixtral 8x7B
MMLU (MCQ in 57 subjects)	69.9%	70.0%	**70.6%**
HellaSwag (10-shot)	**87.1%**	85.5%	86.7%
ARC Challenge (25-shot)	85.1%	85.2%	**85.8%**
WinoGrande (5-shot)	**83.2%**	81.6%	81.2%
MBPP (pass@1)	49.8%	52.2%	**60.7%**
GSM-8K (5-shot)	53.6%	57.1%	**58.4%**
MT Bench (for Instruct Models)	6.86	**8.32**	8.30

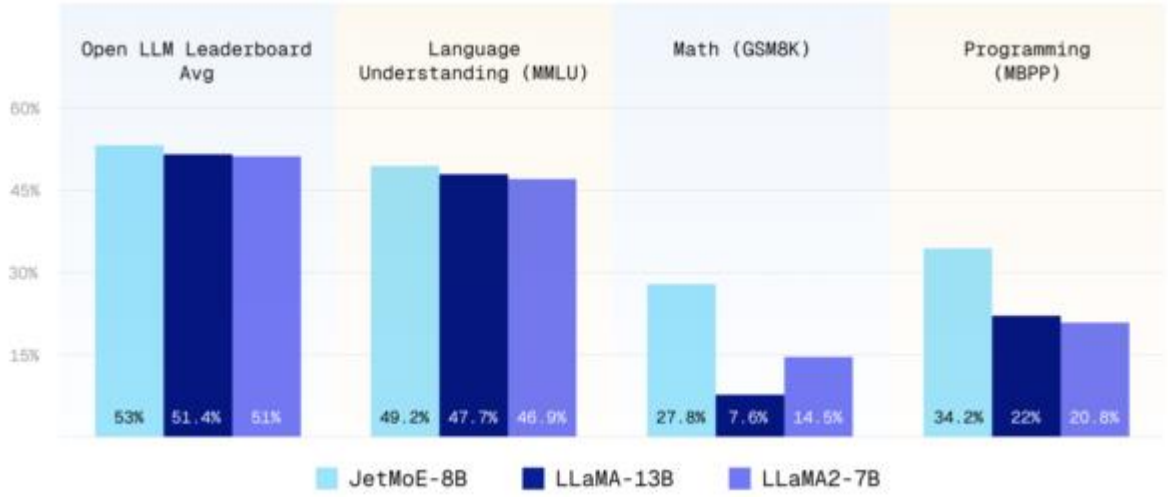
Первая модель LLaMA (Large Language Model/Meta AI) была выпущена компанией Meta (ранее Facebook) в феврале 2023 года. Эта модель была представлена как часть их усилий в области исследований и разработки больших языковых моделей. С тех пор данное семейство продолжает развиваться.



Специализации экспертов не обнаружено

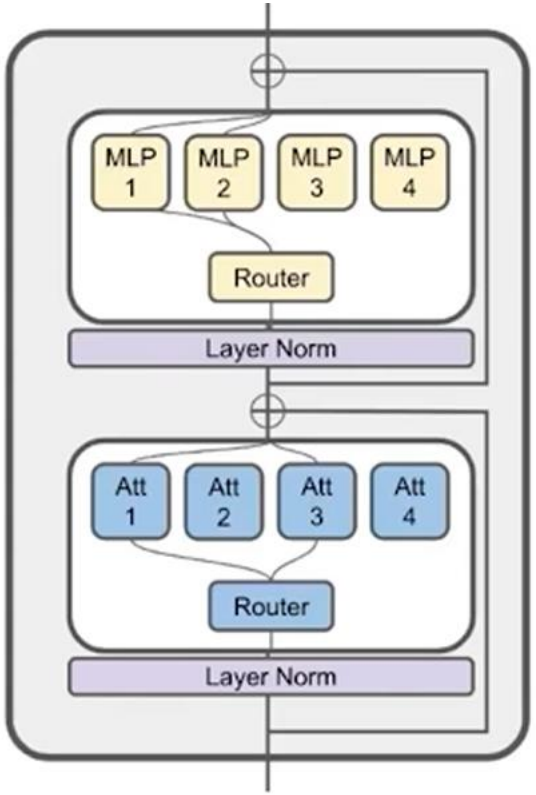
JetMoE

Модель JetMoE-8B, созданная исследователями MIT и Принстонского университета, оказалась лучше LLaMA2-7B. Её тренировка обошлась всего в $0,1 млн, в то время как Meta потратила миллиарды на обучение LLaMA2-7B.



JetMoE

- Комбинация из смеси экспертов и голов внимания.
- 24 блока, 8 экспертов, 2 активных.
- 8 млрд параметров в сумме, ~2 млрд активных.
- Небольшой бюджет: 0.1M долларов.
- 1.3 триллиона высококачественных токенов.

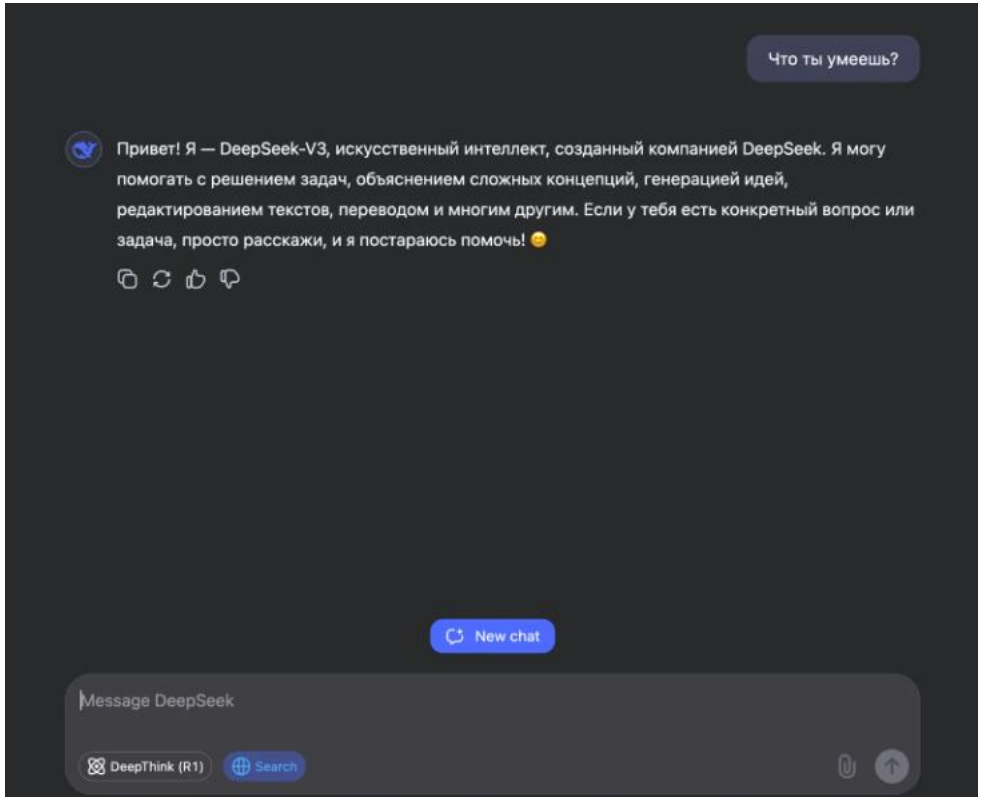


JetMoE

- Идейно подход существует давно и показывал себя достаточно успешно на разных этапах развития отрасли.
- Позволяет ускорить как обучение так и инференс по сравнению с классическими моделями.
- Потребляет существенно большее количество видео памяти.
- Есть проблемы с fine-tuning.
- Nvidia анонсировала новые видеокарты, которые имеют существенно больший размер видео памяти (B200 192GB)

DeepSeek

- DeepSeek — это продвинутая языковая модель искусственного интеллекта, которая работает на основе нейронных сетей и глубинного обучения. Её архитектура основана на технологии трансформеров. Главная особенность трансформеров — это механизм внимания. Он позволяет модели не только анализировать текущее слово в тексте, но и учитывать его связь с остальными словами в предложении.



DeepSeek. Что умеет нейросеть

DeepSeek может выполнять разные задачи, но в основном они связаны с генерацией и обработкой текстовых и числовых данных.
Генерация и понимание текста

- **Создание текстов различных жанров и объёмов.**
- **Анализ и обобщение информации.**

Программирование и разработка кода

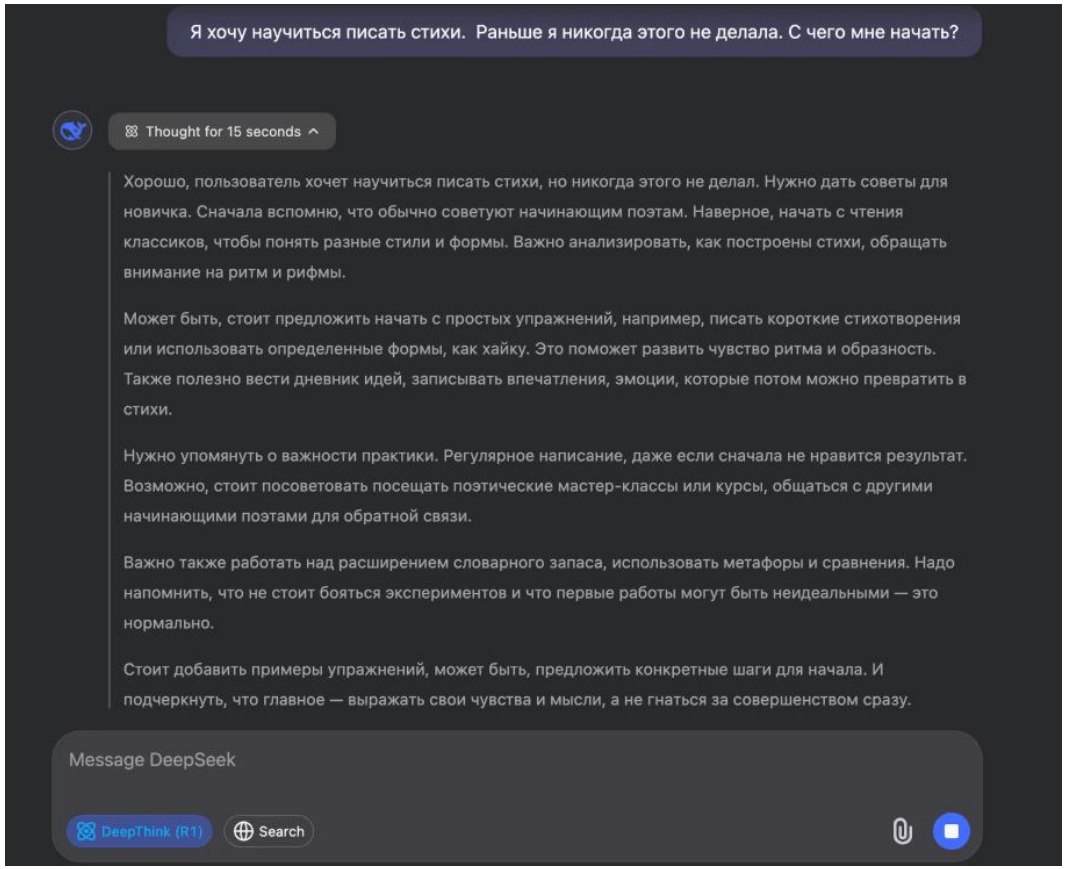
- **Написание и отладка кода.** DeepSeek поддерживает несколько языков программирования: C++, Go, Java, JavaScript, Python, Rust.
- **Интеграция с редакторами кода.** Модель встраивается в разные среды разработки, предоставляет подсказки и автоматизирует задачи.

Поиск и анализ данных

- **Онлайн-поиск информации.** Модель может искать актуальные данные в интернете, предоставлять свежую информацию.
- **Анализ больших данных.** DeepSeek обрабатывает большие объёмы данных, выявляет закономерности.

Логические рассуждения

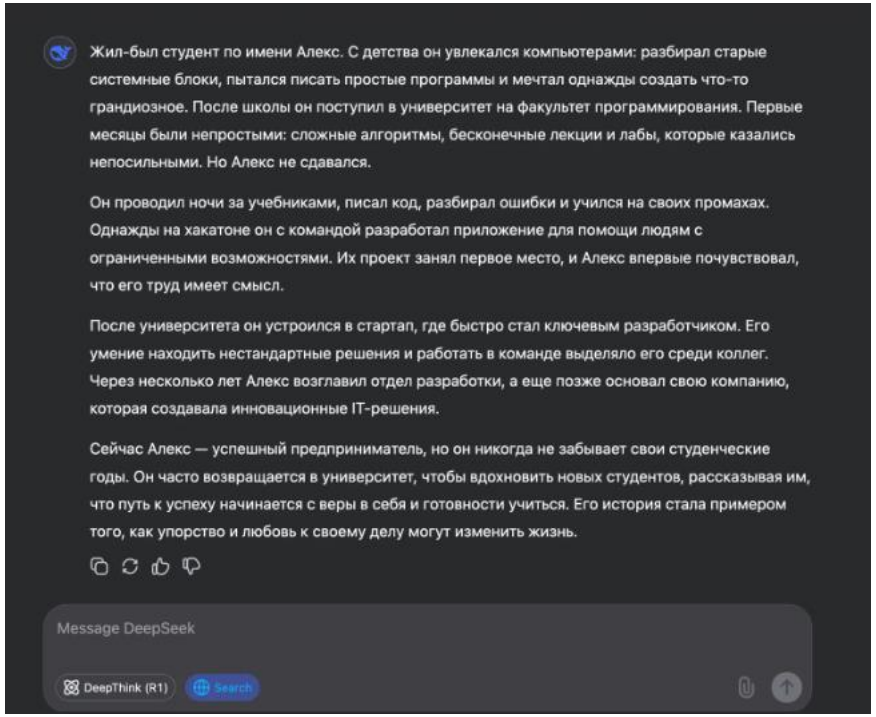
DeepSeek. Что умеет нейросеть



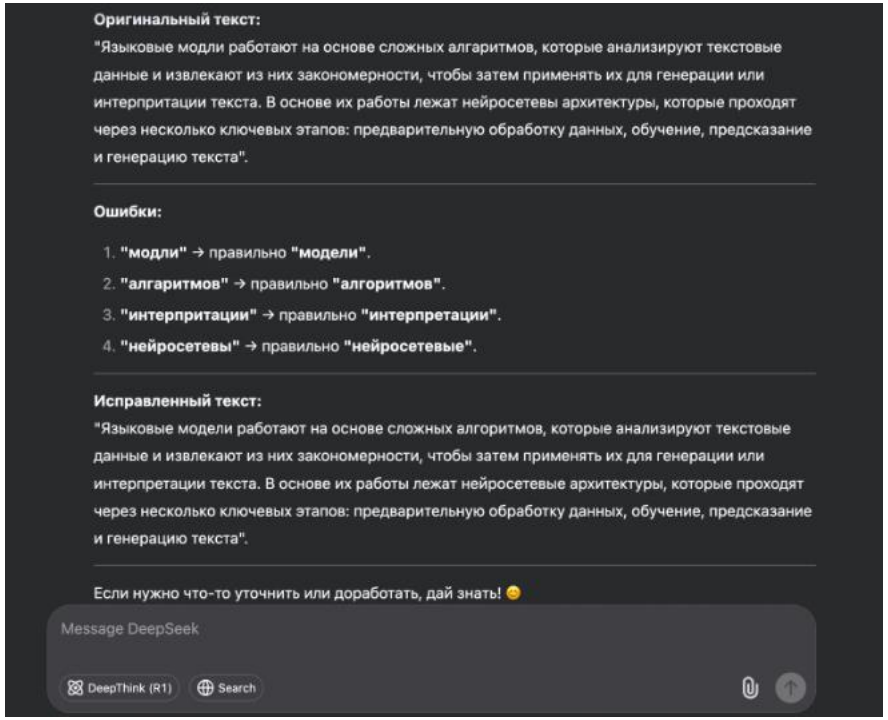
DeepSeek. Преимущества и недостатки

Преимущества	Недостатки
Свободный доступ. Для использования DeepSeek не нужен VPN. Сервис бесплатный, доступен в браузере и в мобильном приложении.	**Язык веб-версии.** Интерфейс в веб-версии DeepSeek может быть на английском и китайском языках (данные на февраль 2025 года), но в мобильной версии — на русском.
Открытый исходный код. Компании могут модифицировать и адаптировать его под свои задачи.	**Сбои в работе.** Иногда из-за большого количества запросов пользователей DeepSeek на время перестаёт генерировать ответы.
Поддержка нескольких языков, в том числе русского.	**Ограниченный функционал.** DeepSeek пока не умеет генерировать изображения и видео. Кроме того, нейросеть не может работать со ссылками и распознавать голосовые запросы.
Высокая производительность. Сравнима с GPT-4 в скорости работы и качестве ответов.	**Не запоминает предыдущий контекст** (в случаях, если работать в нейросети без регистрации).
Можно применять в разных ситуациях. Подходит для генерации текста, программирования, анализа данных и других задач.	
Понимание контекста. Благодаря архитектуре трансформеров DeepSeek хорошо понимает контекст и смысл текста.	
Актуальные данные. DeepSeek, как и многие другие модели, имеет определённую дату обновления. Например, версия V3 — от июля 2024 года. Значит, информация о событиях после этой даты может быть неактуальна. Но спасает функция Search — она подключает к генерации интернет-поиск.	

DeepSeek. Примеры

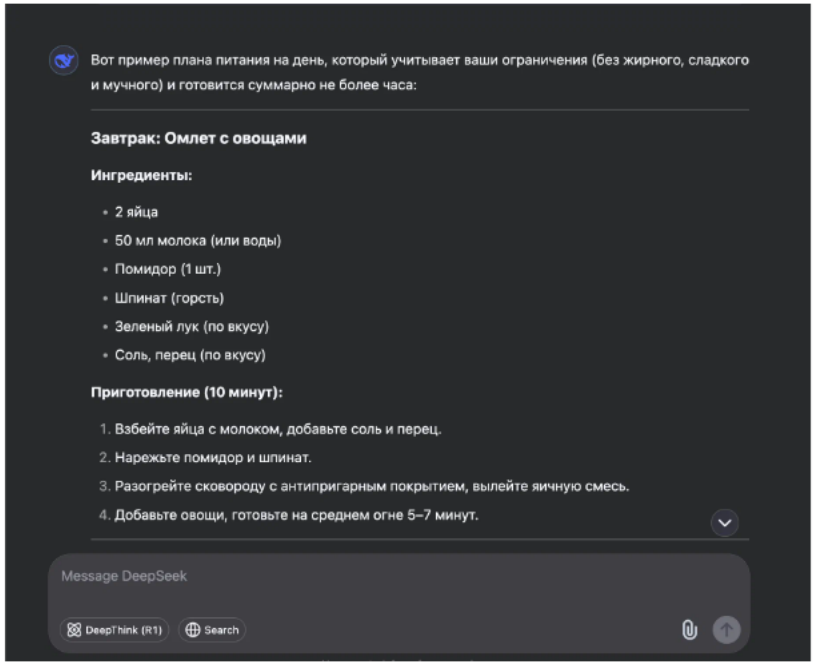


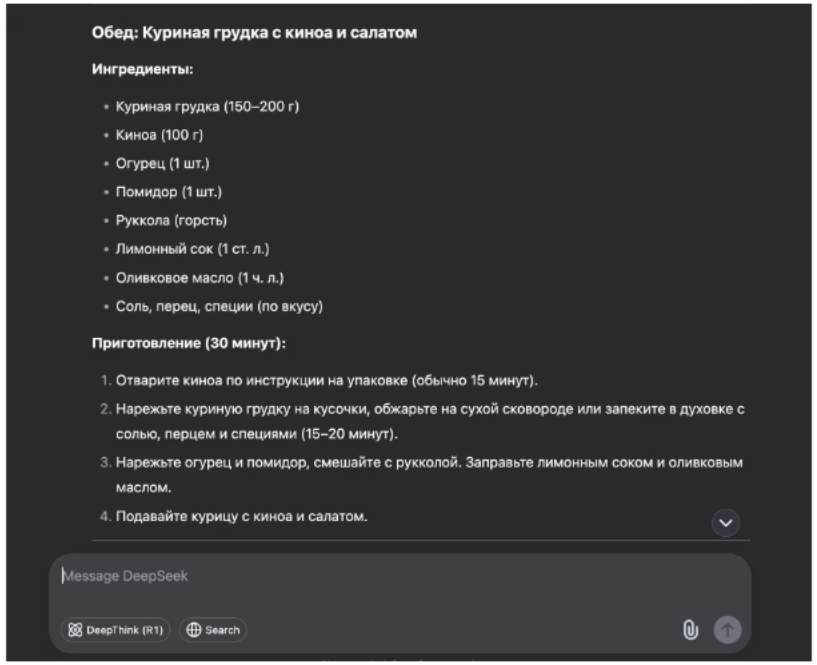
Написать текст



Найти ошибки в тексте

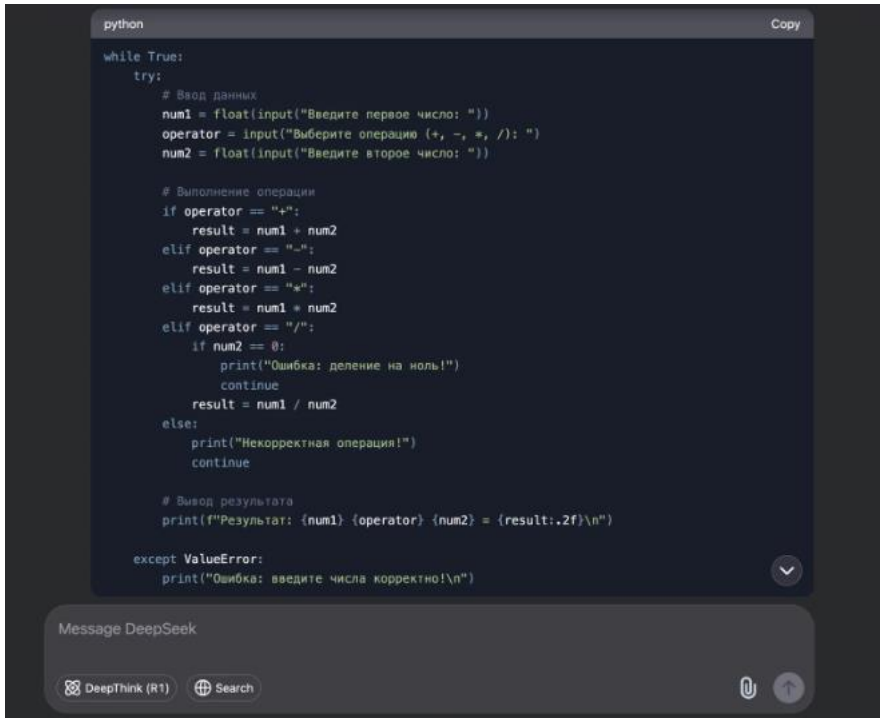
DeepSeek. Примеры





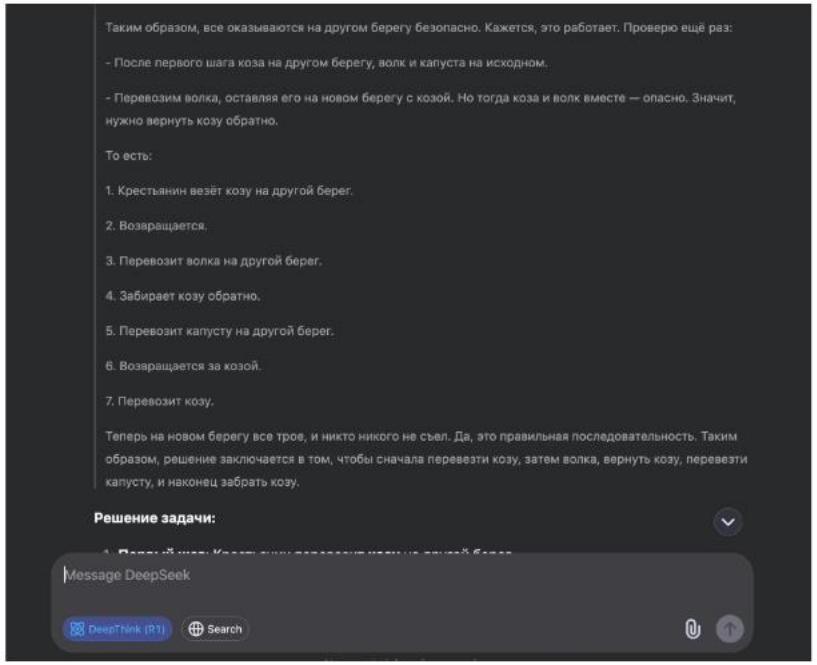
Составить план питания

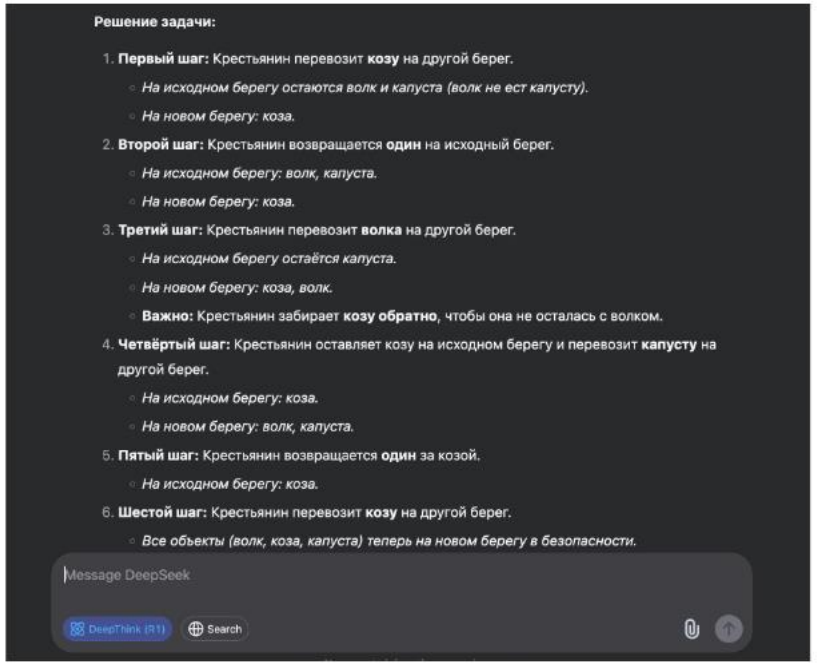
DeepSeek. Примеры



Задание по программированию — написать код простого калькулятора на Python

DeepSeek. Примеры





Задача с логическим рассуждением

Будущее DeepSeek

Развитие в open-source-сообществе. DeepSeek продвигает идею открытого доступа к технологиям ИИ. Её открытый исходный код позволяет разработчикам по всему миру адаптировать и улучшать модель. Так технологии ИИ становятся доступны для малых и средних предприятий, для развивающихся стран.

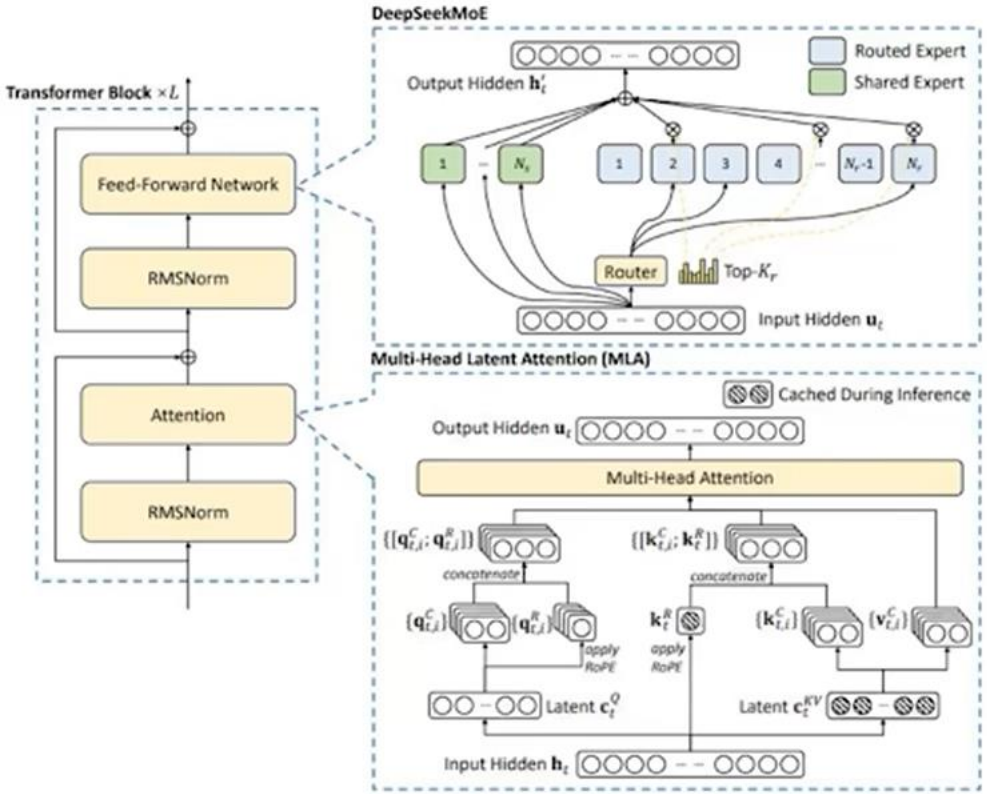
Расширение функционала. В будущем возможны новые функции — генерация аудио и видео.

Конкуренция с OpenAI и Google. Успех DeepSeek показывает, как растёт конкуренция между Китаем и США в сфере ИИ.

Снижение затрат на разработку ИИ в мире. DeepSeek доказала, что создание мощных моделей ИИ не требует огромных финансовых вложений. Его разработка обошлась в 6 млн долларов США — это в десятки раз меньше, чем у конкурентов вроде GPT-4. Поэтому у компаний, которые раньше не могли позволить себе внедрение ИИ, эти возможности теперь появились.

DeepSeek

- MoE в FF слоях
- Эксперты 2 типов: shared и routed
- Более хитрый attention: Multi-Head Latent Attention (MLA)
- Многотокенное прогнозирование (MTP)
- Учили в FP8

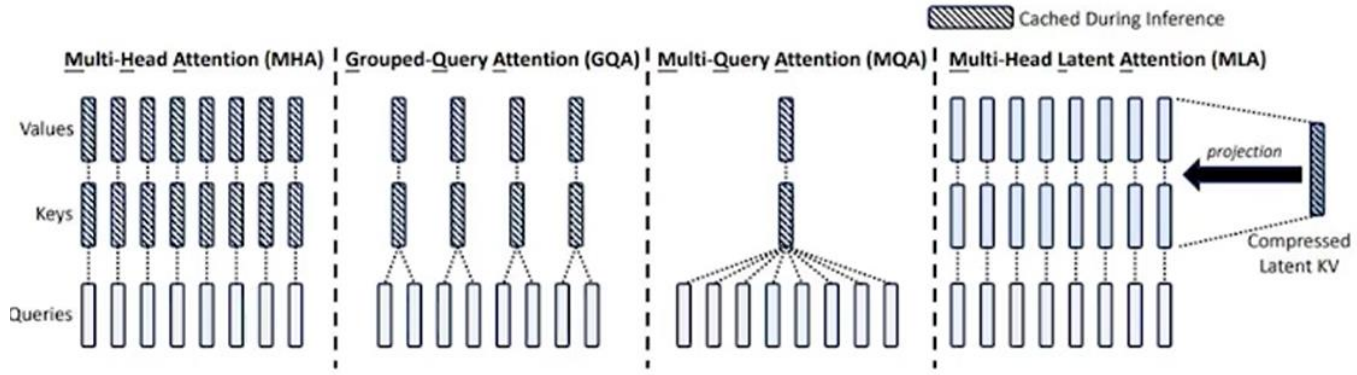


DeepSeek: эксперты, MLA

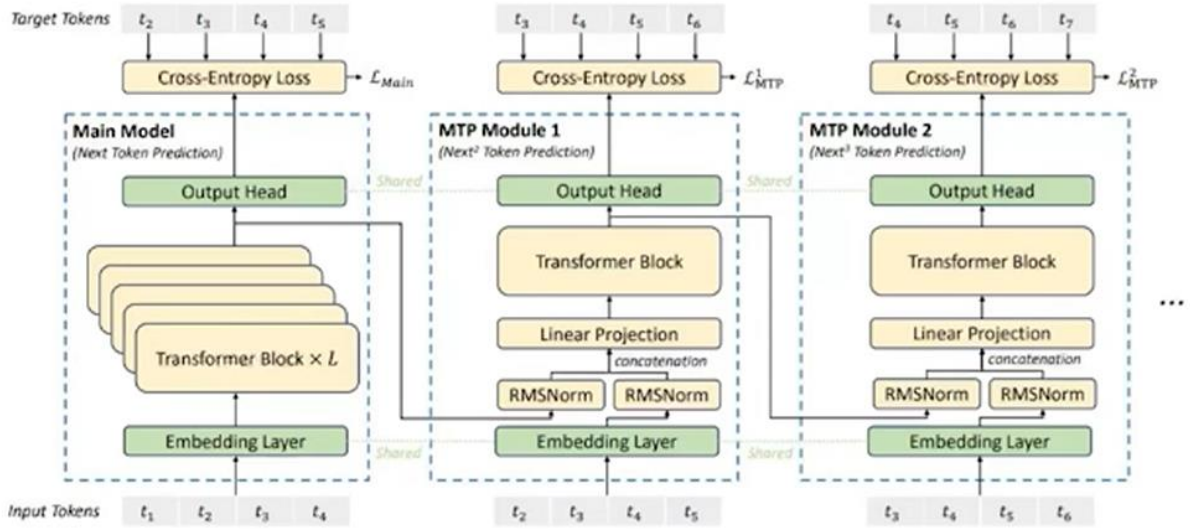
- Fine-grained Experts: разрезаем экспертов на M меньших, роутинг на уровне меньших - можно активировать разное количесэкспертов (более точная специализация - каждый мелкий эксперт фокусируется на узкой области, лучшее распределение нагрузки - больше вариантов для роутинга, гибкость тво мелких экспертов).

- Shared Experts - несколько экспертов активны всегда

- Вместо получения k и v из h напрямую возникает промежуточное звено c. Кэшируем только c. Экономия памяти!



DeepSeek: MTP



$$\mathcal{L}^k_{\text{MTP}} = \text{CrossEntropy}(P^k_{2+k:T+1}, t_{2+k:T+1}) = -\frac{1}{T}\sum_{i=2+k}^{T+1} \log P^k_i[t_i],$$

$$\mathcal{L}_{\text{MTP}} = \frac{\lambda}{D}\sum_{k=1}^{D} \mathcal{L}^k_{\text{MTP}}.$$

DeepSeek: параметры

- 671B параметров, 37B активных
- 1 shared и 256 routed экспертов, 8 активных
- Родные веса - FP8, что означает, что нужны H100+ карты
- 14.8T в обучении

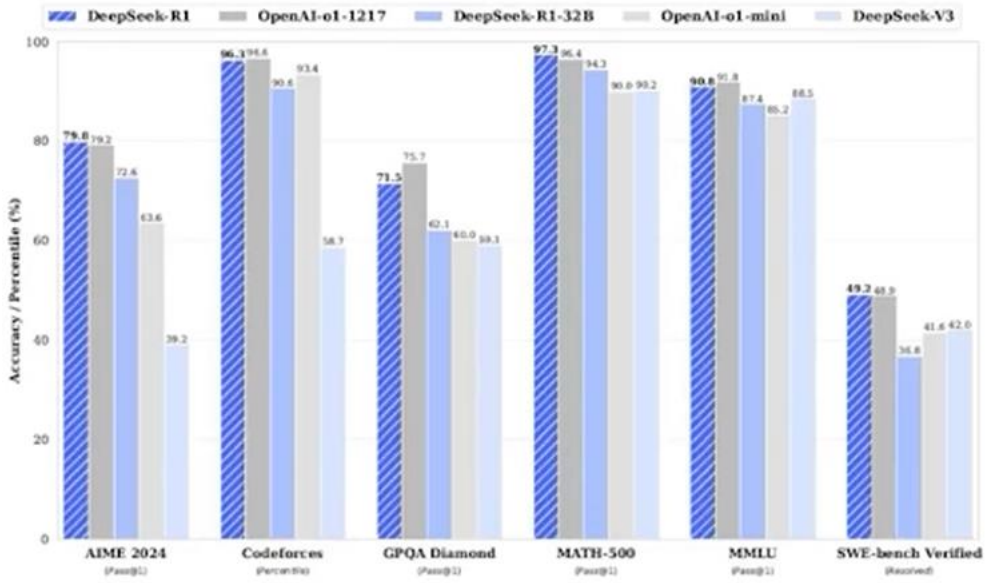
Training Costs	Pre-Training	Context Extension	Post-Training	Total
in H800 GPU Hours	2664K	119K	5K	2788K
in USD	$5.328M	$0.238M	$0.01M	$5.576M

DeepSeek: результаты

	Benchmark (Metric)	# Shots	DeepSeek-V2 Base	Qwen2.5 72B Base	LLaMA-3.1 405B Base	DeepSeek-V3 Base
	Architecture	-	MoE	Dense	Dense	MoE
	# Activated Params	-	21B	72B	405B	37B
	# Total Params	-	236B	72B	405B	671B
English	Pile-test (BPB)	-	0.606	0.638	**0.542**	0.548
	BBH (EM)	3-shot	78.8	79.8	82.9	**87.5**
	MMLU (EM)	5-shot	78.4	85.0	84.4	**87.1**
	MMLU-Redux (EM)	5-shot	75.6	83.2	81.3	**86.2**
	MMLU-Pro (EM)	5-shot	51.4	58.3	52.8	**64.4**
	DROP (F1)	3-shot	80.4	80.6	86.0	**89.0**
	ARC-Easy (EM)	25-shot	97.6	98.4	98.4	**98.9**
	ARC-Challenge (EM)	25-shot	92.2	94.5	**95.3**	**95.3**
	HellaSwag (EM)	10-shot	87.1	84.8	**89.2**	**88.9**
	PIQA (EM)	0-shot	83.9	82.6	**85.9**	84.7
	WinoGrande (EM)	5-shot	**86.3**	82.3	85.2	84.9
	RACE-Middle (EM)	5-shot	73.1	68.1	**74.2**	67.1
	RACE-High (EM)	5-shot	52.6	50.3	**56.8**	51.3
	TriviaQA (EM)	5-shot	80.0	71.9	**82.7**	**82.9**
	NaturalQuestions (EM)	5-shot	38.6	33.2	**41.5**	40.0
	AGIEval (EM)	0-shot	57.5	75.8	60.6	**79.6**
Code	HumanEval (Pass@1)	0-shot	43.3	53.0	54.9	**65.2**
	MBPP (Pass@1)	3-shot	65.0	72.6	68.4	**75.4**
	LiveCodeBench-Base (Pass@1)	3-shot	11.6	12.9	15.5	**19.4**
	CRUXEval-I (EM)	2-shot	52.5	59.1	58.5	**67.3**
	CRUXEval-O (EM)	2-shot	49.8	59.9	59.9	**69.8**
Math	GSM8K (EM)	8-shot	81.6	88.3	83.5	**89.3**
	MATH (EM)	4-shot	43.4	54.4	49.0	**61.6**
	MGSM (EM)	8-shot	63.6	76.2	69.9	**79.8**
	CMath (EM)	3-shot	78.7	84.5	77.3	**90.7**
Chinese	CLUEWSC (EM)	5-shot	82.0	82.5	**83.0**	**82.7**
	C-Eval (EM)	5-shot	81.4	89.2	72.5	**90.1**
	CMMLU (EM)	5-shot	84.0	**89.5**	73.7	88.8
	CMRC (EM)	1-shot	**77.4**	75.8	76.0	76.3
	C3 (EM)	0-shot	77.4	76.7	**79.7**	78.6
	CCPM (EM)	0-shot	**93.0**	88.5	78.6	92.0
Multilingual	MMMLU-non-English (EM)	5-shot	64.0	74.8	73.8	**79.4**

DeepSeek R1

- Популярность началась именно с R1 (хотя V3 вышла несколько раньше)
- Является одной из первых reasoning открытых моделей
- Выпустили много разных distill версий
- Соперник для O1



DeepSeek R1

Model	AIME 2024		MATH-500	GPQA Diamond	LiveCode Bench	CodeForces
	pass@1	cons@64	pass@1	pass@1	pass@1	rating
GPT-4o-0513	9.3	13.4	74.6	49.9	32.9	759
Claude-3.5-Sonnet-1022	16.0	26.7	78.3	65.0	38.9	717
OpenAI-o1-mini	63.6	80.0	90.0	60.0	53.8	**1820**
QwQ-32B-Preview	50.0	60.0	90.6	54.5	41.9	1316
DeepSeek-R1-Distill-Qwen-1.5B	28.9	52.7	83.9	33.8	16.9	954
DeepSeek-R1-Distill-Qwen-7B	55.5	83.3	92.8	49.1	37.6	1189
DeepSeek-R1-Distill-Qwen-14B	69.7	80.0	93.9	59.1	53.1	1481
DeepSeek-R1-Distill-Qwen-32B	**72.6**	83.3	94.3	62.1	57.2	1691
DeepSeek-R1-Distill-Llama-8B	50.4	80.0	89.1	49.0	39.6	1205
DeepSeek-R1-Distill-Llama-70B	70.0	**86.7**	**94.5**	**65.2**	**57.5**	1633

Qwen3

Qwen3 создала Alibaba. Это компания, которая управляет AliExpress. Несмотря на новизну — а представили ее меньше года назад — нейросеть получилась очень умная. Во-первых, она поддерживает 119 языков, включая русский. А, во-вторых, обучалась на огромном массиве данных из 36 триллионов токенов.

Ее учили не только на текстах, но и на программном коде, научных статьях, задачах на логику и колоссальном объеме синтетических данных.

Qwen3

Компактные модели нужны для простых задач. Например, **Qwen3–0.6B и Qwen3–1.7B** могут запускаться просто на обычных видеокартах, не требуют серверов и отлично подходят для мобильных приложений, чат-ботов, локальных помощников. Такие модели быстро отвечают на короткие вопросы, справляются с переводами, базовой генерацией текста и не перегружают систему.

Если задача сложнее, подойдут **Qwen3–4B или Qwen3–8B**. Они уже способны держать длинный контекст и работать с многоязычными интерфейсами, но потребуют и больших вычислительных мощностей.

Models	Layers	Heads (Q / KV)	Tie Embedding	Context Length
Qwen3-0.6B	28	16 / 8	Yes	32K
Qwen3-1.7B	28	16 / 8	Yes	32K
Qwen3-4B	36	32 / 8	Yes	32K
Qwen3-8B	36	32 / 8	No	128K
Qwen3-14B	40	40 / 8	No	128K
Qwen3-32B	64	64 / 8	No	128K

Models	Layers	Heads (Q / KV)	# Experts (Total / Activated)	Context Length
Qwen3-30B-A3B	48	32 / 4	128 / 8	128K
Qwen3-235B-A22B	94	64 / 4	128 / 8	128K

Qwen3

В Qwen3 есть два режима: thinking и non-thinking. **В thinking mode модель рассуждает поэтапно**. Она анализирует задачу, разбивает её на логические шаги и только потом формирует ответ.

В **non-thinking режиме Qwen3** не показывает, как она думает, соответственно, тратит меньше ресурсов и работает максимально быстро. Здесь нет лишних объяснений — только короткий ответ по делу.

Qwen3

Модель	Архитектура	Параметров (млрд)	Контекст, токенов	Языки	Режимы мышления	Работа с файлами	Открытая лицензия
Qwen3-235B (онлайн)	MoE	235 (22 активных)	128 000	119	Да	Да	Да
GPT-4o (ChatGPT)	Dense	175	128 000	56	Нет	Нет	Нет
Gemini 2.5 Pro	MoE	340 (MoE)	128 000	100+	Нет	Нет	Нет
DeepSeek-R1	MoE	670 (37 активных)	128 000	100+	Нет	Нет	Да

Qwen3. Пример

Перевод статьи из интернета

Бывший генеральный директор компании Ookla, которая создала популярное приложение Speedtest, недавно представил новое мобильное приложение под названием **NetSure** , предназначенное для оценки **надежности интернет-соединения** , а не только скорости загрузки и передачи данных.

В отличие от большинства других инструментов тестирования интернета, которые фокусируются исключительно на пропускной способности, NetSure стремится дать пользователям более полное представление о том, **насколько стабильным и предсказуемым является их соединение** . Это особенно важно в эпоху удаленной работы, онлайн-обучения и потоковой передачи видео.

🧠 Что делает NetSure особенным?

Ookla известна своим приложением **Speedtest** , которое используется миллиардами людей по всему миру для проверки скорости интернета. Однако Speedtest измеряет лишь базовые параметры: загрузку, выгрузку и пинг.

NetSure же добавляет к этим метрикам **анализ стабильности соединения** . Приложение может определить, как часто соединение теряется или становится медленным, а также показывает **статистику задержек (латентности)** и **колебаний скорости** за последние 7 дней.

Пользователь получает:

- Среднюю скорость
- Минимальную и максимальную скорость
- Время безотказной работы

Qwen3. Пример

Написать код

2. Persistence.swift

swift

```swift
import Foundation
import CoreData

public class PersistenceController {
    public static let shared = PersistenceContainer()

    public static var preview: PersistenceContainer = {
        let result = PersistenceContainer(inMemory: true)
        let viewContext = result.container.viewContext
        do {
            try viewContext.save()
        } catch {
            let nsError = error as NSError
            fatalError("Unresolved error $nsError), $nsError.userInfo)")
        }
        return result
    }()
}

class PersistenceContainer: ObservableObject {
    let container: NSPersistentCloudKitContainer
    var viewContext: NSManagedObjectCo xt { container.viewContext }

```

Qwen3. Пример

Здравоохранение города Москвы
Городская поликлиника №
ОГРН
г. Москва
Тел.

Справка

Дана Иванову Ивану Ивановичу 01.01.1980 г.р.

В том что он(а) «01» января 20 15 г. прошел (ла) медицинское обследование, необходимое для участия в соревнованиях по теннису
(указать вид спорта)

Соматическая патология **выявлена/Не выявлена**
(нужное подчеркнуть)

ЭКГ без патологий

Заключение противопоказаний для участия в соревнованиях по теннису не выявлено

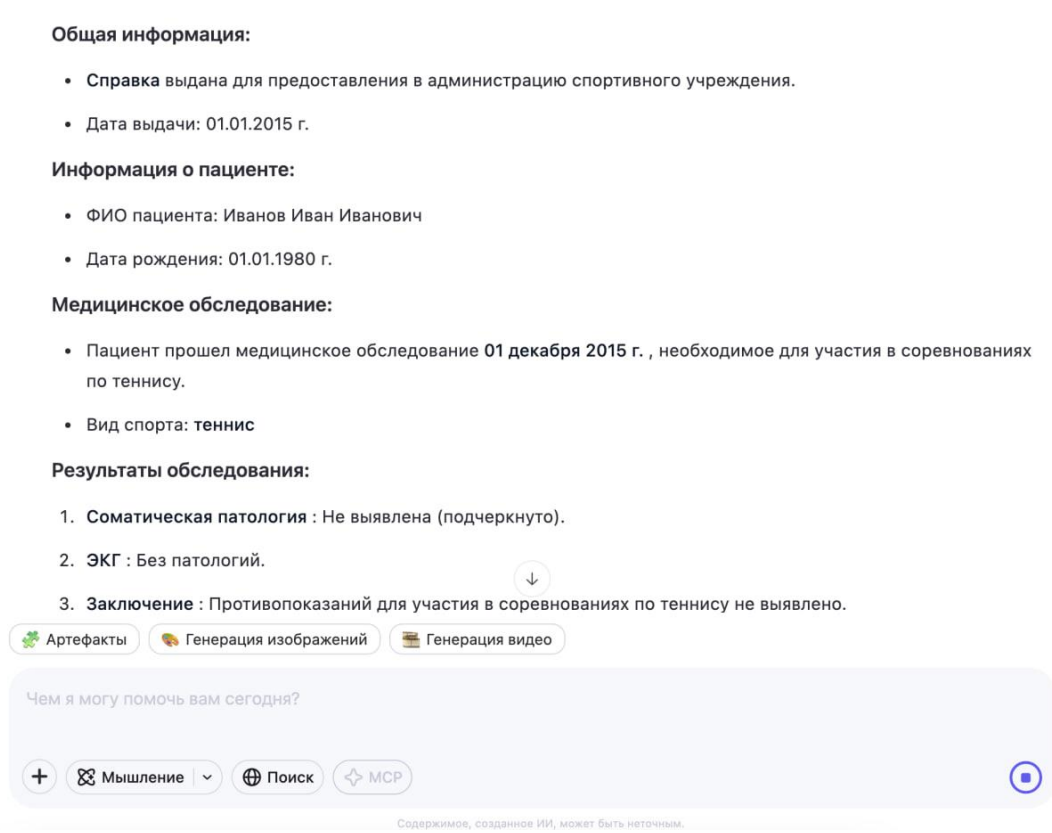
Справка дана для предоставления в администрацию спортивного учреждения.

ДЛЯ СПРАВОК

М.П.

Врач

ВРАЧ ТЕРАПЕВТ



Общая информация:

- **Справка** выдана для предоставления в администрацию спортивного учреждения.
- Дата выдачи: 01.01.2015 г.

Информация о пациенте:

- ФИО пациента: Иванов Иван Иванович
- Дата рождения: 01.01.1980 г.

Медицинское обследование:

- Пациент прошел медицинское обследование **01 декабря 2015 г.** , необходимое для участия в соревнованиях по теннису.
- Вид спорта: **теннис**

Результаты обследования:

1. **Соматическая патология** : Не выявлена (подчеркнуто).
2. **ЭКГ** : Без патологий.
3. **Заключение** : Противопоказаний для участия в соревнованиях по теннису не выявлено.

Артефакты | Генерация изображений | Генерация видео

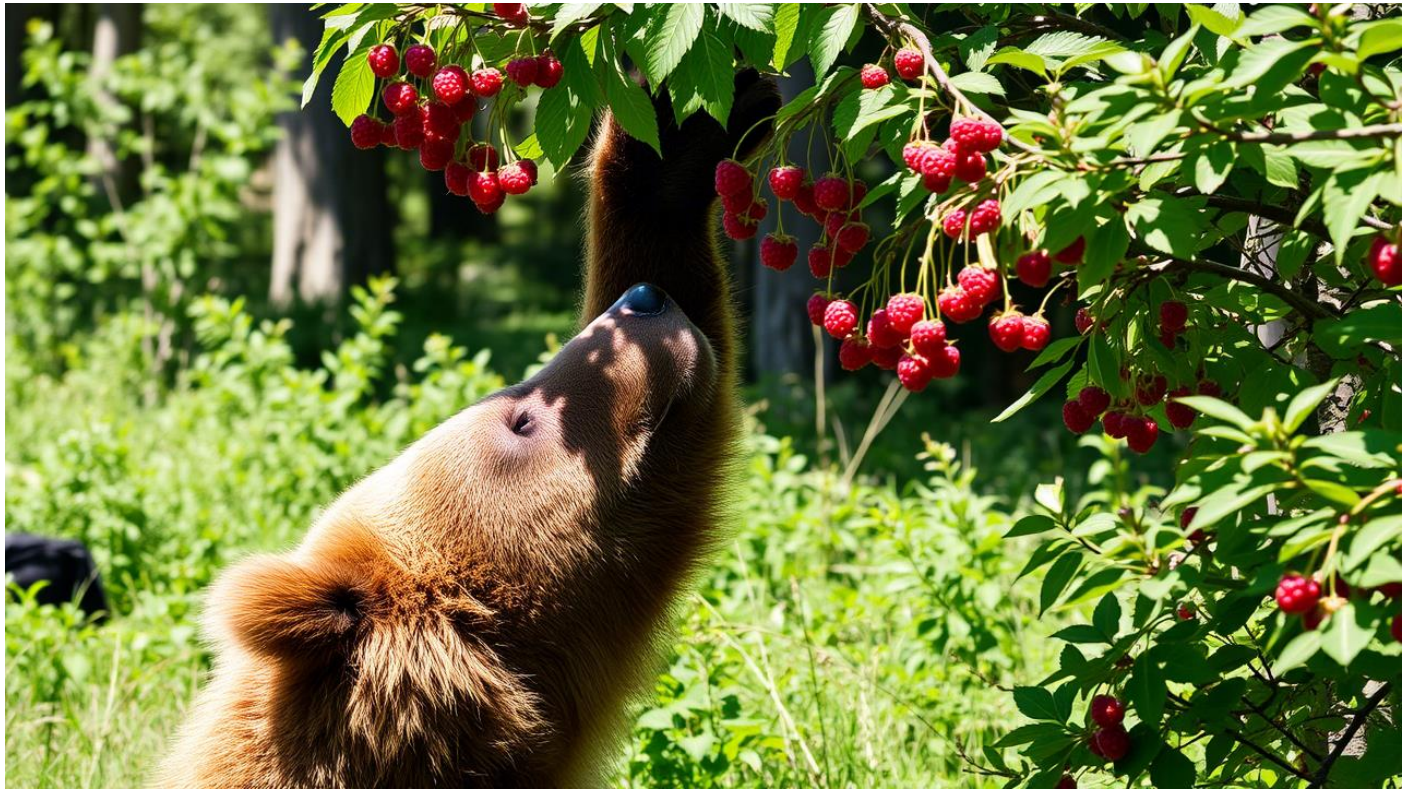
Чем я могу помочь вам сегодня?

Мышление | Поиск | MCP

Содержимое, созданное ИИ, может быть неточным.

Выделить рукописный текст

Qwen3. Пример



Генерация изображений

Qwen3

- Шаг 1: обычный pre-training
- 30 триллионов токенов, 119 языков
- 4096 длина последовательности
- Шаг 2: ризонинг
- 5 триллионов высококачественных токенов, содержащих STEM, код, рассуждения, синтетические данные
- Шаг 3: Длинный контекст
- Данные содержащие примеры до длиной 32768 токенов
- 100+ миллиардов токенов

Qwen3

	Qwen2.5-72B Base	Qwen2.5-Plus Base	Llama-4-Maverick Base	DeepSeek-V3 Base	Qwen3-235B-A22B Base
Architecture	Dense	MoE	MoE	MoE	MoE
# Total Params	72B	271B	402B	671B	235B
# Activated Params	72B	37B	17B	37B	22B
General Tasks					
MMLU	86.06	85.02	85.16	87.19	**87.81**
MMLU-Redux	83.91	82.69	84.05	86.14	**87.40**
MMLU-Pro	58.07	63.52	63.91	59.84	**68.18**
SuperGPQA	36.20	37.18	40.85	41.53	**44.06**
BBH	86.30	85.60	83.62	86.22	**88.87**
Math & STEM Tasks					
GPQA	45.88	41.92	43.94	41.92	**47.47**
GSM8K	91.50	91.89	87.72	87.57	**94.39**
MATH	62.12	62.78	63.32	62.62	**71.84**
Coding Tasks					
EvalPlus	65.93	61.43	68.38	63.75	**77.60**
MultiPL-E	58.70	62.16	57.28	62.26	**65.94**
MBPP	76.00	74.60	75.40	74.20	**81.40**
CRUX-O	66.20	68.50	77.00	76.60	**79.00**
Multilingual Tasks					
MGSM	82.40	82.21	79.69	82.68	**83.53**
MMMLU	84.40	83.49	83.09	85.88	**86.70**
INCLUDE	69.05	66.97	73.47	**75.17**	73.46

	Gemma-3-12B Base	Qwen2.5-14B Base	Qwen2.5-32B Base	Qwen2.5-Turbo Base	Qwen3-14B Base	Qwen3-30B-A3B Base
Architecture	Dense	Dense	Dense	MoE	Dense	MoE
# Total Params	12B	14B	32B	42B	14B	30B
# Activated Params	12B	14B	32B	6B	14B	3B
General Tasks						
MMLU	73.87	79.66	**83.32**	79.50	81.05	81.38
MMLU-Redux	70.70	76.64	**81.97**	77.11	79.88	81.17
MMLU-Pro	44.91	51.16	55.10	55.60	61.03	**61.49**
SuperGPQA	24.61	30.68	33.55	31.19	34.27	**35.72**
BBH	74.28	78.18	**84.48**	76.10	81.07	81.54
Math & STEM Tasks						
GPQA	31.31	32.83	**47.97**	41.41	39.90	43.94
GSM8K	78.01	90.22	**92.87**	88.32	92.49	91.81
MATH	44.43	55.64	57.70	55.60	**62.02**	59.04
Coding Tasks						
EvalPlus	52.65	60.70	66.25	61.23	**72.23**	71.45
MultiPL-E	43.03	54.79	58.30	53.24	61.69	**66.53**
MBPP	60.60	69.00	73.60	67.60	73.40	**74.40**
CRUX-O	52.00	61.10	67.80	60.20	**68.60**	67.20
Multilingual Tasks						
MGSM	64.35	74.68	78.12	70.45	**79.20**	79.11
MMMLU	72.50	78.34	**82.40**	79.76	79.69	81.46
INCLUDE	63.34	60.26	64.35	59.25	64.55	**67.00**

Пост-трейнинг

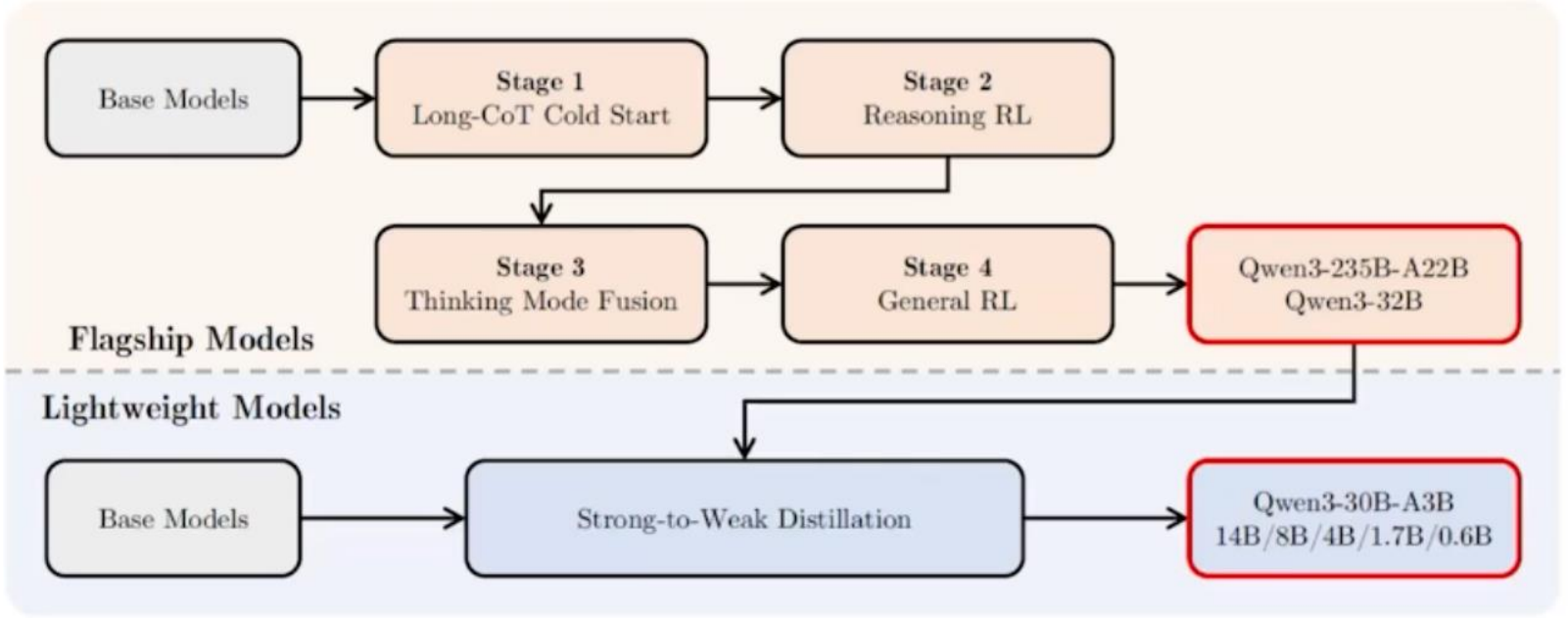


Figure 1: Post-training pipeline of the Qwen3 series models.

Пост-трейнинг

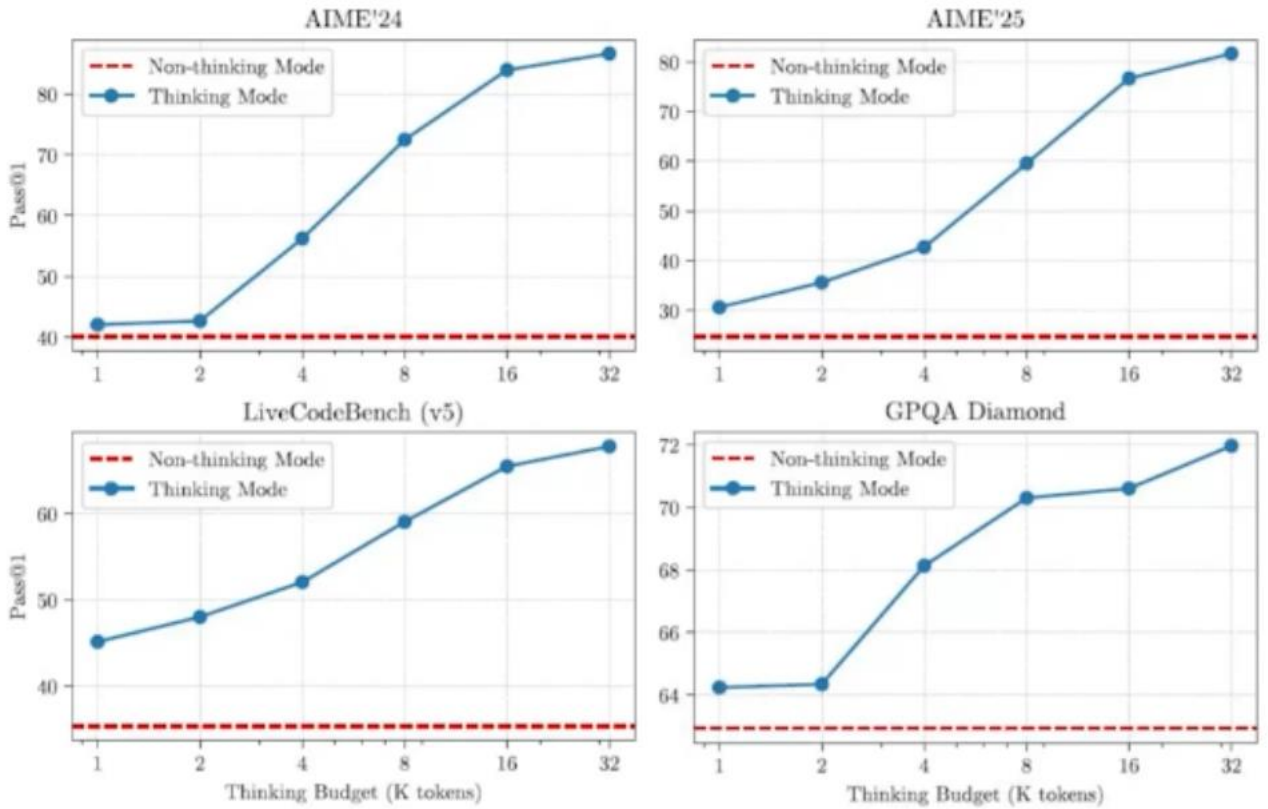
- гибридный ризонинг - модели могут как размышлять, так и сразу отвечать
- /no_think токен в конец user content части - модель не будет размышлять
- можно в рамках одного диалога использовать оба варианта
- дистилляция от сильной к слабой
- сначала для топовых моделей (32B и 235B-A22B) провели многоступенчатую схему обучения
- затем дистиллировали в меньшие модели в 2 шага: дообучение на генерациях и дистилляция

Влияние шагов дообучения на качество

Table 22: Performance of Qwen3-32B after Reasoning RL (Stage 2), Thinking Mode Fusion (Stage 3), and General RL (Stage 4). Benchmarks with * are in-house datasets.

		Stage 2 Reasoning RL	Stage 3 Thinking Mode Fusion		Stage 4 General RL	
	Benchmark	Thinking	Thinking	Non-Thinking	Thinking	Non-Thinking
General Tasks	LiveBench 2024-11-25	68.6	70.9 +2.3	57.1	74.9 +4.0	59.8 +2.8
	Arena-Hard	86.8	89.4 +2.6	88.5	93.8 +4.4	92.8 +4.3
	CounterFactQA*	50.4	61.3 +10.9	64.3	68.1 +6.8	66.4 +2.1
Instruction & Format Following	IFEval strict prompt	73.0	78.4 +5.4	78.4	85.0 +6.6	83.2 +4.8
	Multi-IF	61.4	64.6 +3.2	65.2	73.0 +8.4	70.7 +5.5
	LengthCtrl*	62.6	70.6 +8.0	84.9	73.5 +2.9	87.3 +2.4
	ThinkFollow*	-	88.7		98.9 +10.2	
Agent	BFCL v3	69.0	68.4 -0.6	61.5	70.3 +1.9	63.0 +1.5
	ToolUse*	63.3	70.4 +7.1	73.2	85.5 +15.1	86.5 +13.3
Knowledge & STEM	MMLU-Redux	91.4	91.0 -0.4	86.7	90.9 -0.1	85.7 -1.0
	GPQA-Diamond	68.8	69.0 +0.2	50.4	68.4 -0.6	54.6 +4.3
Math & Coding	AIME'24	83.8	81.9 -1.9	28.5	81.4 -0.5	31.0 +2.5
	LiveCodeBench v5	68.4	67.2 -1.2	31.1	65.7 -1.5	31.3 +0.2

Влияние количества размышлений на качество



Дистилляция

Table 21: Comparison of reinforcement learning and on-policy distillation on Qwen3-8B. Numbers in parentheses indicate pass@64 scores.

Method	AIME'24	AIME'25	MATH500	LiveCodeBench v5	MMLU -Redux	GPQA -Diamond	GPU Hours
Off-policy Distillation	55.0 (90.0)	42.8 (83.3)	92.4	42.0	86.4	55.6	-
+ Reinforcement Learning	67.6 (90.0)	55.5 (83.3)	94.8	52.9	86.9	61.3	17,920
+ On-policy Distillation	**74.4 (93.3)**	**65.5 (86.7)**	**97.0**	**60.3**	**88.3**	**63.3**	1,800

Off-policy - модель дообучалась на генерациях

On-policy - модель дообучалась дополнительно на логитах бОльшей модели!

Метрики на русском языке

Table 32: **Benchmark scores for language: Russian (ru)**. The highest and second-best scores are shown in **bold** and underlined, respectively.

	Model	Multi-IF	INCLUDE	MT-AIME24	PolyMath	Average
Thinking Mode	Gemini2.5-Pro	**68.1**	**80.4**	70.0	52.3	67.7
	QwQ-32B	61.2	73.2	76.7	43.6	63.7
	Qwen3-235B-A22B	62.2	**80.4**	**80.0**	**53.1**	**68.9**
	Qwen3-32B	62.5	73.2	63.3	46.5	61.4
	Qwen3-30B-A3B	60.7	76.8	73.3	45.4	64.0
	Qwen3-14B	63.6	**80.4**	66.7	46.4	64.3
	Qwen3-8B	62.9	69.6	63.3	37.7	58.4
	Qwen3-4B	52.8	69.6	56.7	36.6	53.9
	Qwen3-1.7B	37.8	46.4	20.0	22.8	31.8
	Qwen3-0.6B	26.4	46.4	3.3	7.0	20.8
Non-thinking Mode	GPT-4o-2024-1120	52.0	**80.4**	20.0	13.7	41.5
	Gemma-3-27b-IT	57.3	71.4	23.3	21.6	43.4
	Qwen2.5-72B-Instruct	54.1	67.9	20.0	13.3	38.8
	Qwen3-235B-A22B	56.7	75.0	**40.0**	**26.1**	**49.4**
	Qwen3-32B	58.6	71.4	30.0	23.3	45.8
	Qwen3-30B-A3B	58.0	73.2	30.0	21.1	45.6
	Qwen3-14B	**60.3**	71.4	26.7	24.2	45.6
	Qwen3-8B	59.3	58.9	20.0	22.8	40.2
	Qwen3-4B	46.1	58.9	13.3	17.8	34.0
	Qwen3-1.7B	34.8	41.1	3.3	13.2	23.1
	Qwen3-0.6B	25.5	46.4	0.0	5.8	19.4