

Как мы делаем Нейропоиск

Смирнов Тимофей

Почему поиск нужен LLM

Актуальность

Тимофей
где сейчас холоднее, в москве или лондоне

Нейро
В Москве холоднее, чем в Лондоне.
В Москве 15 ноября 2024 года температура в 01:29 составляет +1 °C, пасмурно, влажность 81%, ветер 3 м/с, вероятность осадков 0%. В Лондоне в тот же день днём +9 °C, пасмурно, влажность 75%, ветер 3 м/с, вероятность осадков 0%. [yandex.ru](#) • [yandex.ru](#)

Климат Москвы — умеренно-континентальный, с снежными и морозными зимами. В Лондоне климат умеренный, зимы прохладные, но не морозные, даже ночью в зимнее время года редко бывает ниже 0 °C. [otvet.mail.ru](#) • [dzen.ru](#)

Найти в Помощнике

Ответ сформирован YandexGPT на основе текстов выбранных сайтов. В нём могут быть неточности.

Сходили в API Яндекс.погоды для Москвы и для Лондона и сравнили погоду

президент сша

поиск с нейро картинки видео карты товары переводчик все

Нейро
На основе источников, возможны неточности

Действующий президент США — Джо Байден. [kommersant.ru](#) • [whitehouse.gov](#)
5 ноября 2024 года в стране прошли выборы президента, на которых победил Дональд Трамп. Он получил 312 голосов выборщиков по всей стране при необходимых 270. Его соперница от демократов Камала Харрис получила 226 голосов. [kommersant.ru](#)

В должность президента США Трамп вступит 20 января 2025 года. Вице-президентом страны станет сенатор от Огайо Джей Ди Вэнс. [kommersant.ru](#) • [RBC.ru](#)

13 ноября 2024 года Байден принял в Овальном кабинете Белого дома избранного президента Трампа. [kommersant.ru](#)

Уточните вопрос

Меньше галлюцинации в сложных фактах

- Бенчмарк галлюцинаций лм от openai
- accuracy_given_attempted - точность среди ответов где модель дала ответ
- $f1 = f1(accuracy_given_attempted, total_accuracy)$

model name	f1
neuro-prod	0.824
gpt-4o-2024-08-06	0.401
gpt-4o-mini-2024-07-18	0.095

<https://cdn.openai.com/papers/simpleqa.pdf>

Как оценить LLM претрейн



Human Eval

Haystack index/quote

DOCBasket

MMLU

HellaSwag

COQA

LogiQA

WINOGRANDE

Natural Questions

ARC_Challenge

DaNetQA

MIPT_CULT

COPA

MUSR

MBPP

FLORES

LAMBADA

CRAG

ARC

PIQA

GSM8k

BoolQ

Squad

BookWorm

Math

RACE

MMLU Pro

ASDiv

DROP

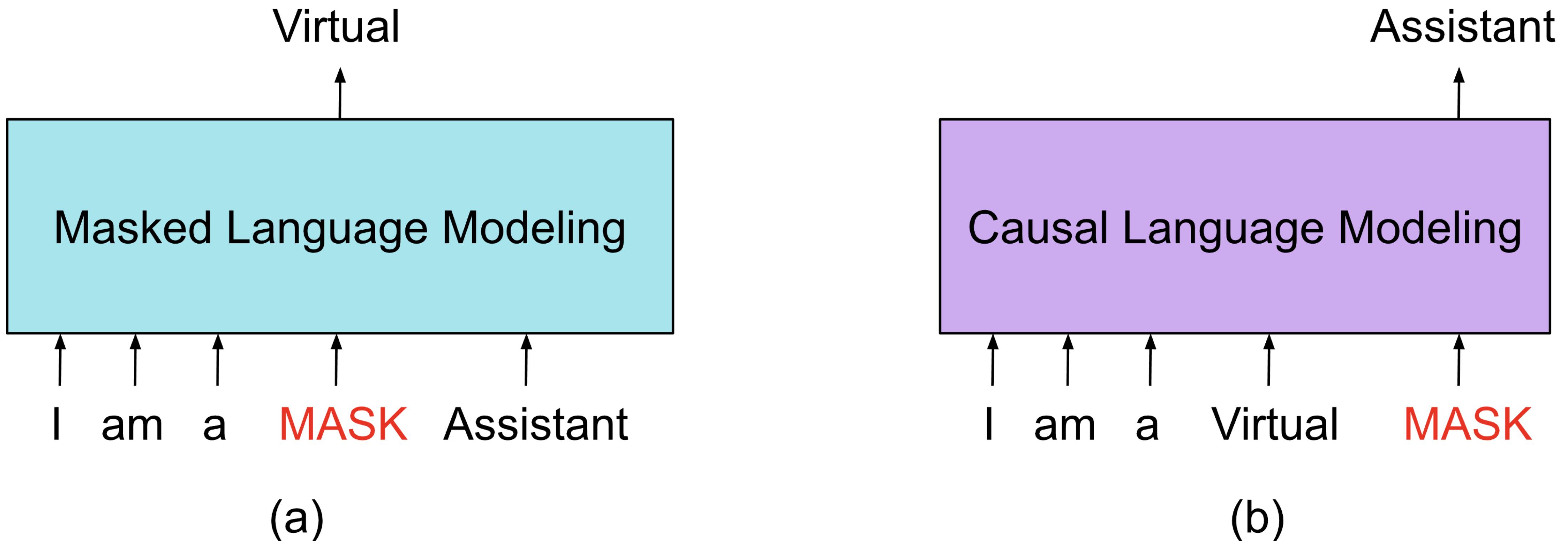
CommonSenseQA

Как оценить LLM претрейн

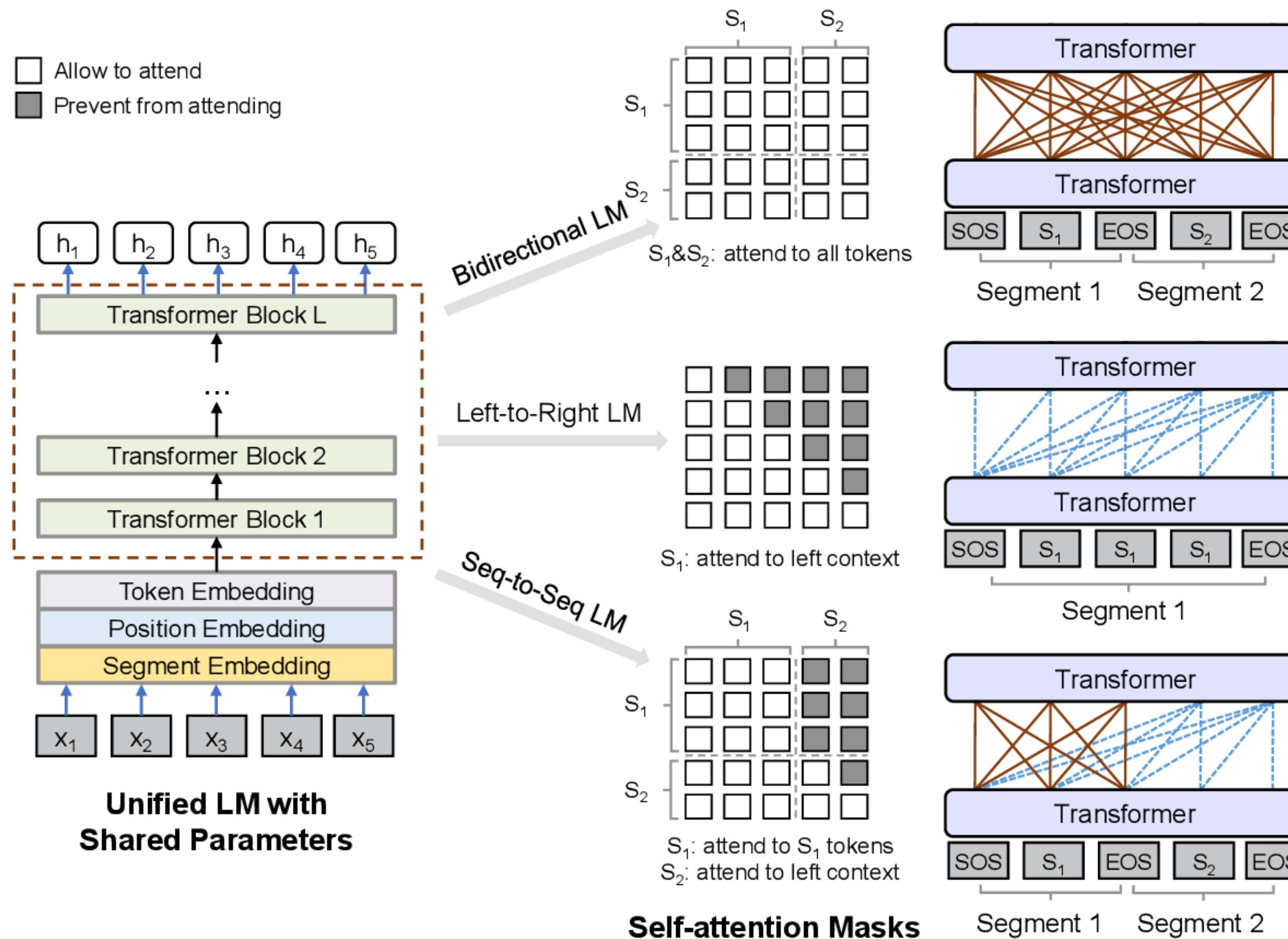
Категориазция

- Common sense - winograd/CommonsenseQA/HellaSwag/...
- Code - HumanEval/MBPP/...
- Exams - MMLU/MMLU Pro/ЕГЭ/ОГЭ/AP/...
- Format - json/feval
- Knowledge - arc/natural questions/simpleqa
- Reading - quality/haystack in the needle/...
- Math - gsm8k/asdiv/math/...
- Reasoning - LogiQA/MUSR/CRAG

Bert/GPT/PrefixLM Reminder

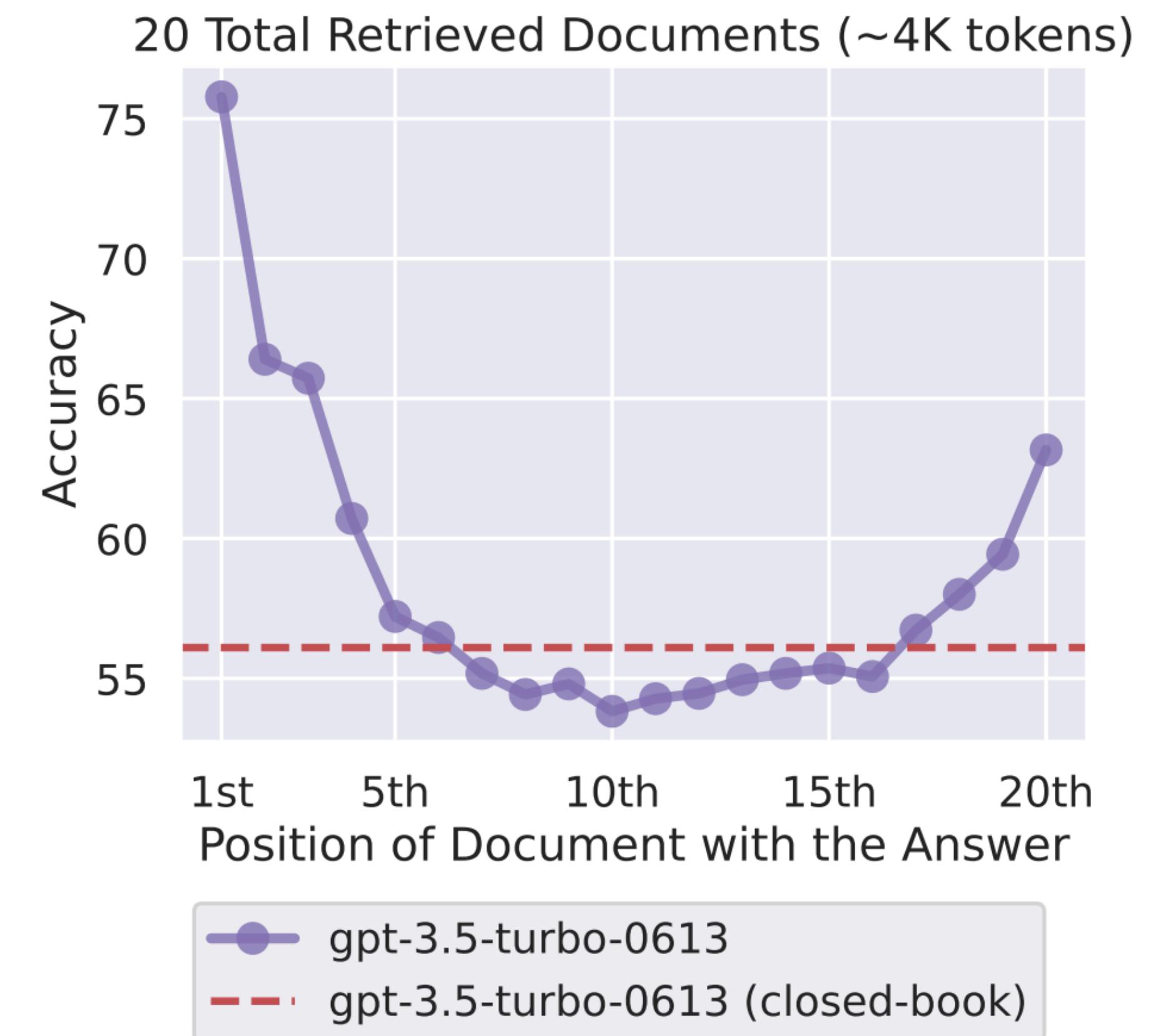


Bert/GPT/PrefixLM Reminder



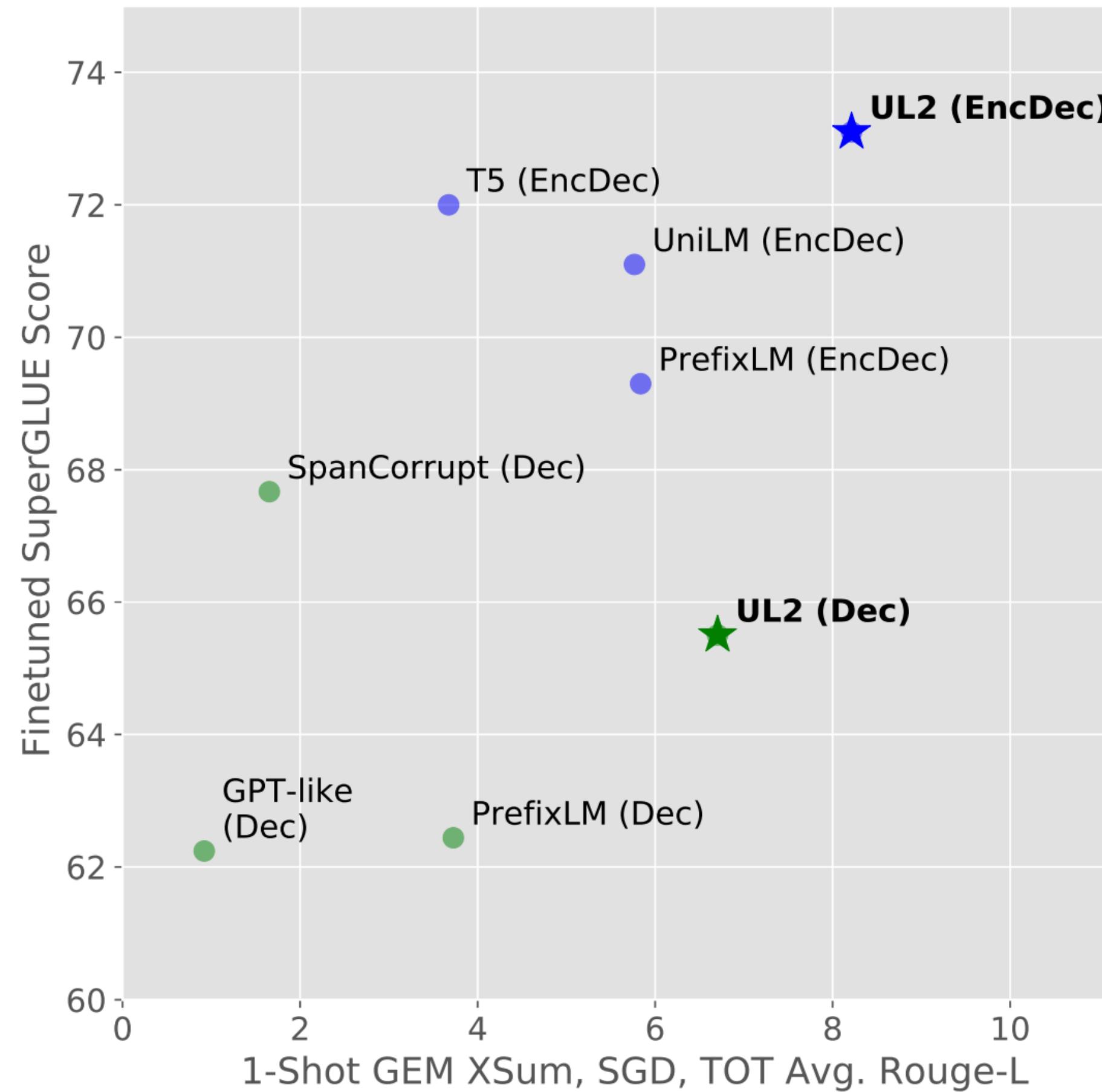
Lost-in-the-middle problem

- Отбирают на вопрос 20 документов и меняют позицию правильного документа.
- В случае если правильный документ первый или последний LLM работает сильно лучше
- На момент середины 23-го все LLM страдали этой проблемой



Unifying Language Learning Paradigms

Баланс качества генеративных и классификационных задач



Unifying Language Learning Paradigms

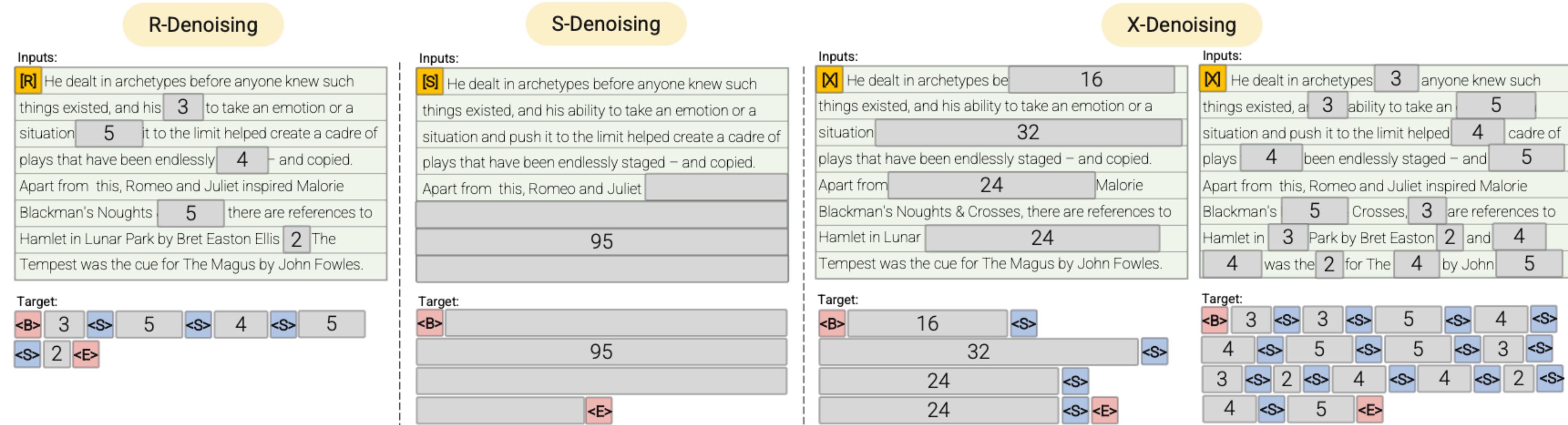


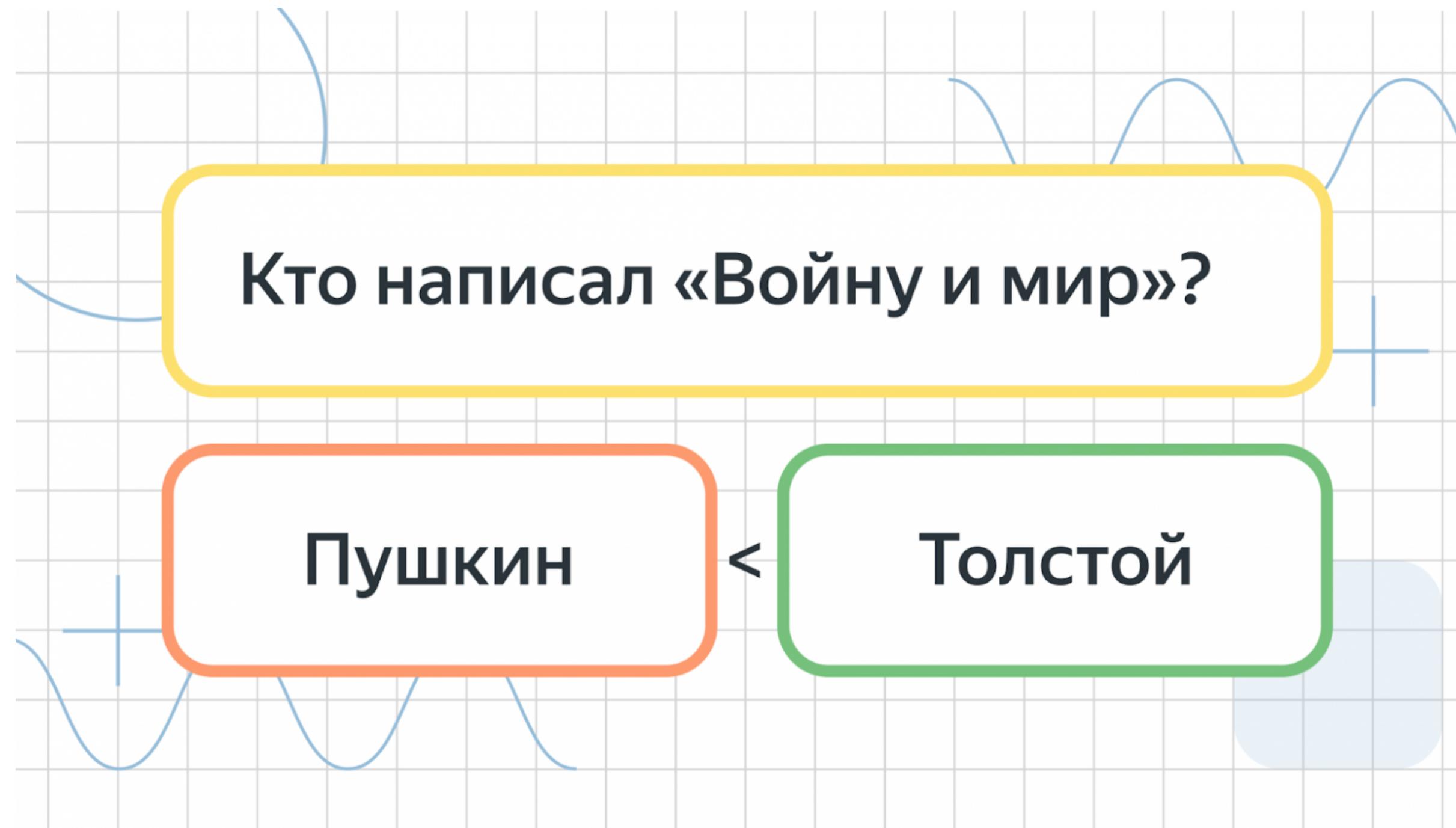
Figure 3: Mixture of denoisers for training UL2. Greyed out rectangles are masked tokens that are shifted to ‘targets’ for prediction.

Reword

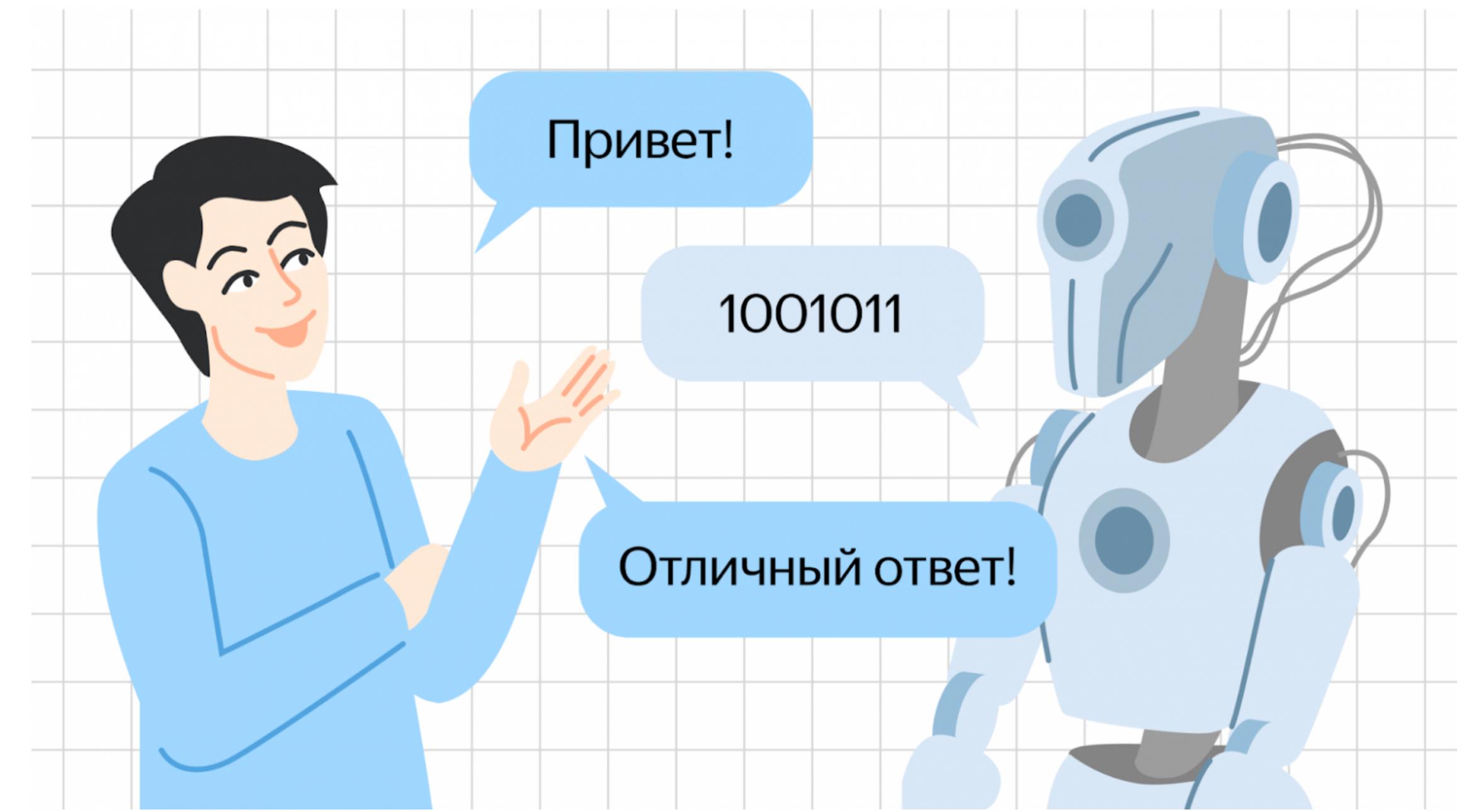
Reword

Как оценивать качество

Pairwise оценка



Pointwise оценка



Reword

Почему pointwise плох

- Pointwise оценки менее гибкие: новые неожиданные генерации случайно попадают в какой-то класс: sbs - надо переучить, pointwise - надо править инструкцию
- Когда в pointwise много аспектов тяжело контролировать качество
- Тем не менее в каких-то понятных задачах pointwise вполне себе работает, примеры: подтверждность/фактологичность/орфография

Reword

Bradley-Terry

- R - реворд модель оценивающая “силу” ответа
- x - инстракт
- y₁ - правый ответ
- y₂ - левый ответ

$$P(y_1 > y_2 | x) = \frac{\exp(R(x, y_1))}{\exp(R(x, y_1)) + \exp(R(x, y_2))}$$

$$L(x, y_1, y_2) = - (I(y_1 > y_2) \cdot \log(p(y_1 > y_2 | x)) + I(y_2 > y_1) \cdot \log(p(y_2 > y_1 | x)))$$

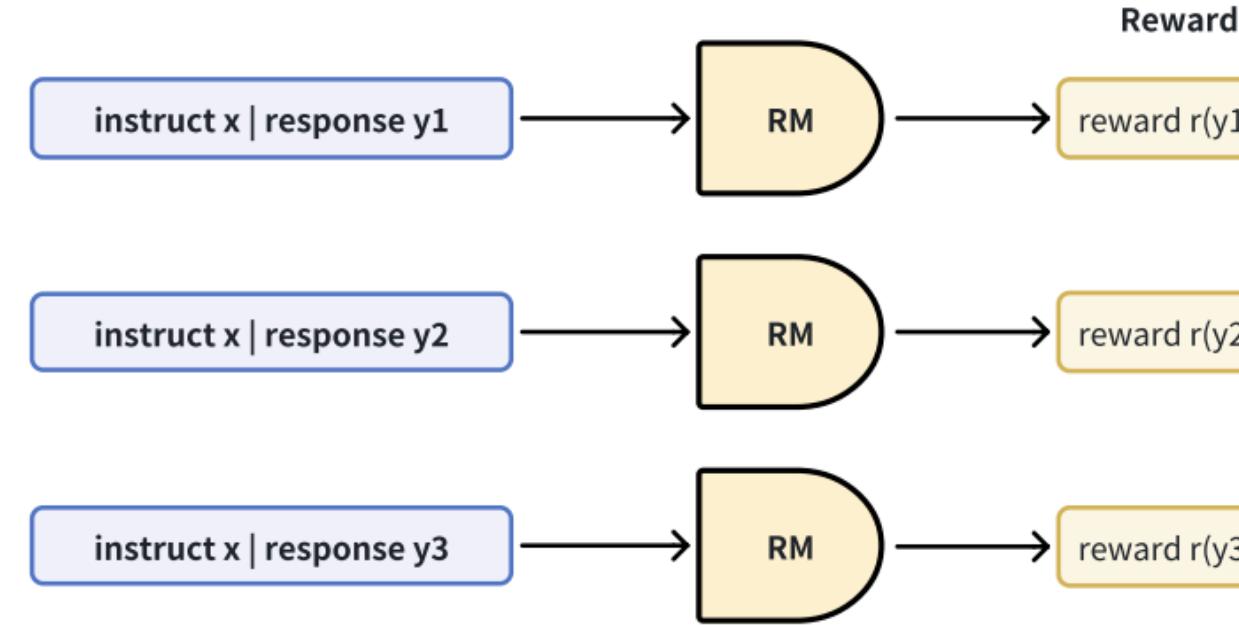
Reword

Оба ответа за раз

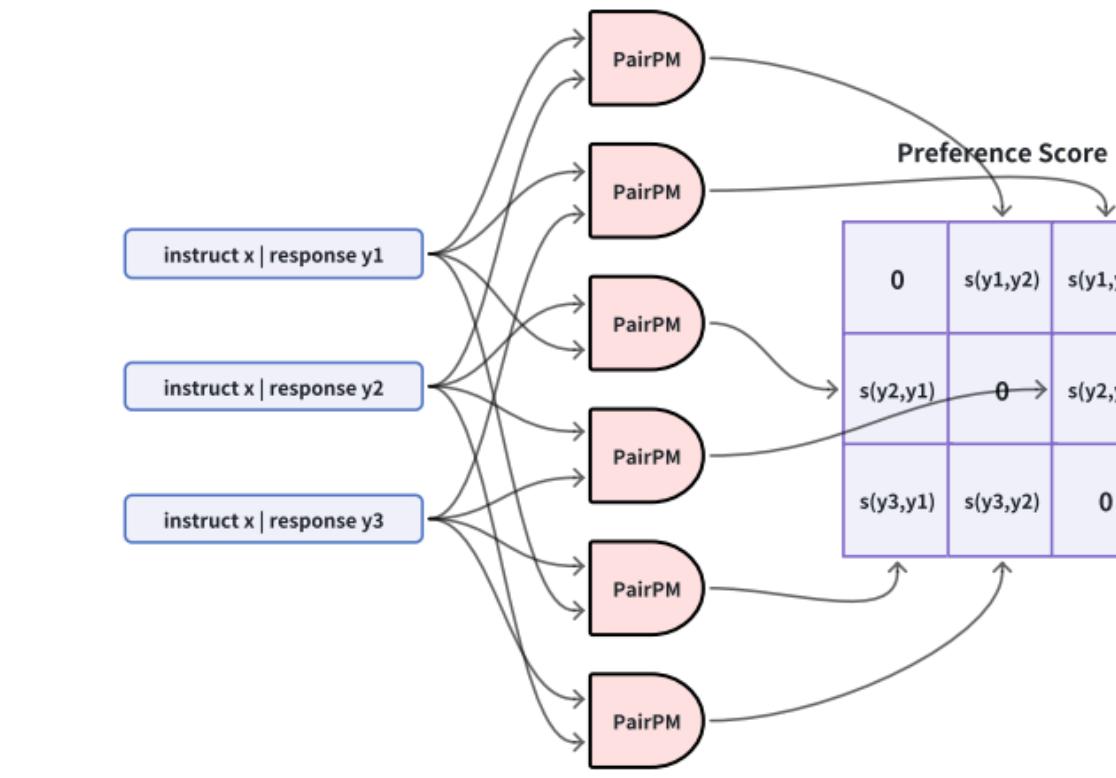
- Модель Бредли-Тэрри не разрешает циклы. Невозможно чтобы было $A > B, B > C$ и при этом $C > A$
- При этом существует Парадокс Кондорсе
- Формально теперь не $R(x, y_1)$ а $R(x, y_1, y_2)$
- Модель получается антисимметрична

Reword

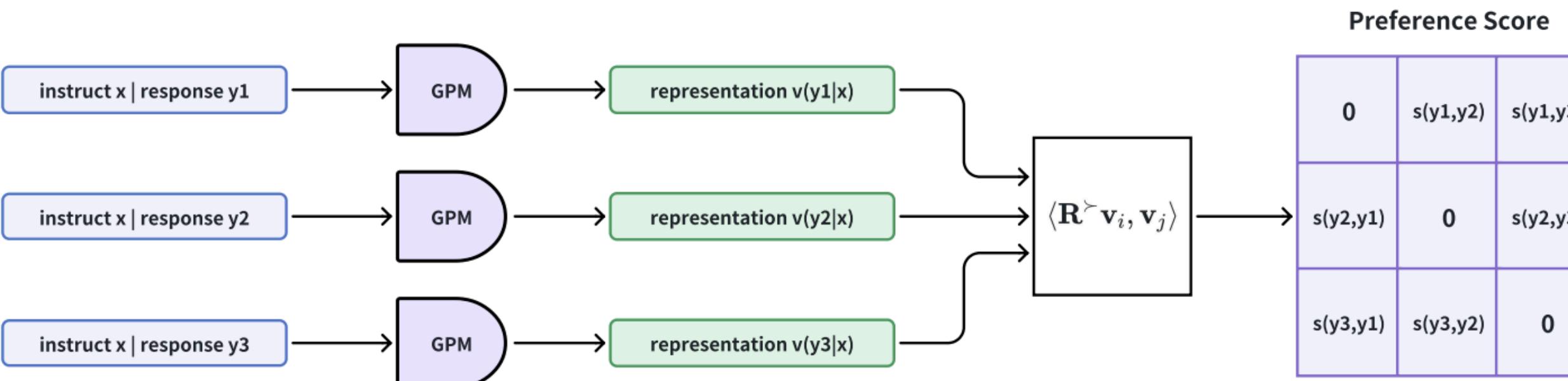
General Preference Modeling with Preference Representations



(a) Bradley-Terry (BT) reward model



(b) PairPM



(c) General Preference representation model (GPM)

Reword

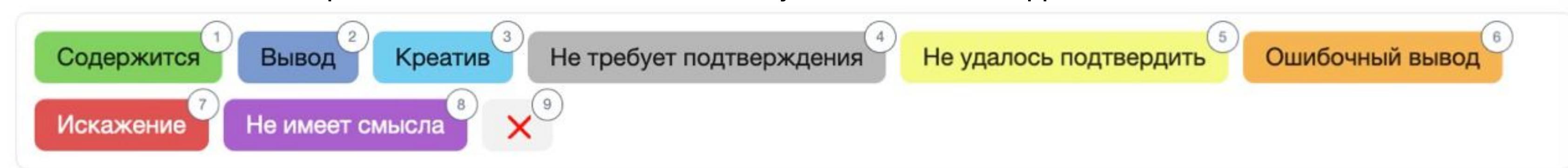
Аспекты

- **Полезность.** Ответ даёт нужную информацию в удобном для восприятия виде.
- **Безопасность.** В ответе нет очевидно вредной или опасной рекомендации.
- **Подтверждённость.** Если необходимо, ответ подтверждён ссылками на источники информации.
- **Компетентность.** Ответ не содержит явных ошибок, структурирован, грамотен, логично и последовательно изложен
- **Main** - итоговый аспект качества
- На каждый аспект pairwise разметка + pointwise

Reword

Подтвержденностъ

Запрос - Сколько обычно значений у многозначного фразеологизма?



Фразеологизмы, как и слова, могут быть однозначными и многозначными [1] [2] [3] .

Большинство фразеологизмов, как и слов, однозначны [5] . Они имеют одно значение [2] .

Многозначных фразеологизмов меньше [1] [2] [3] .

Однозначные фразеологизмы совпадают по значению со словами, а многозначные — с выражениями [5] .

У многозначных фразеологизмов есть несколько значений [1] [2] [3] . Например, «считать ворон» может означать:

- «быть рассеянным» [1] ;
- «бездельничать» [1] .

SFT

SFT TO

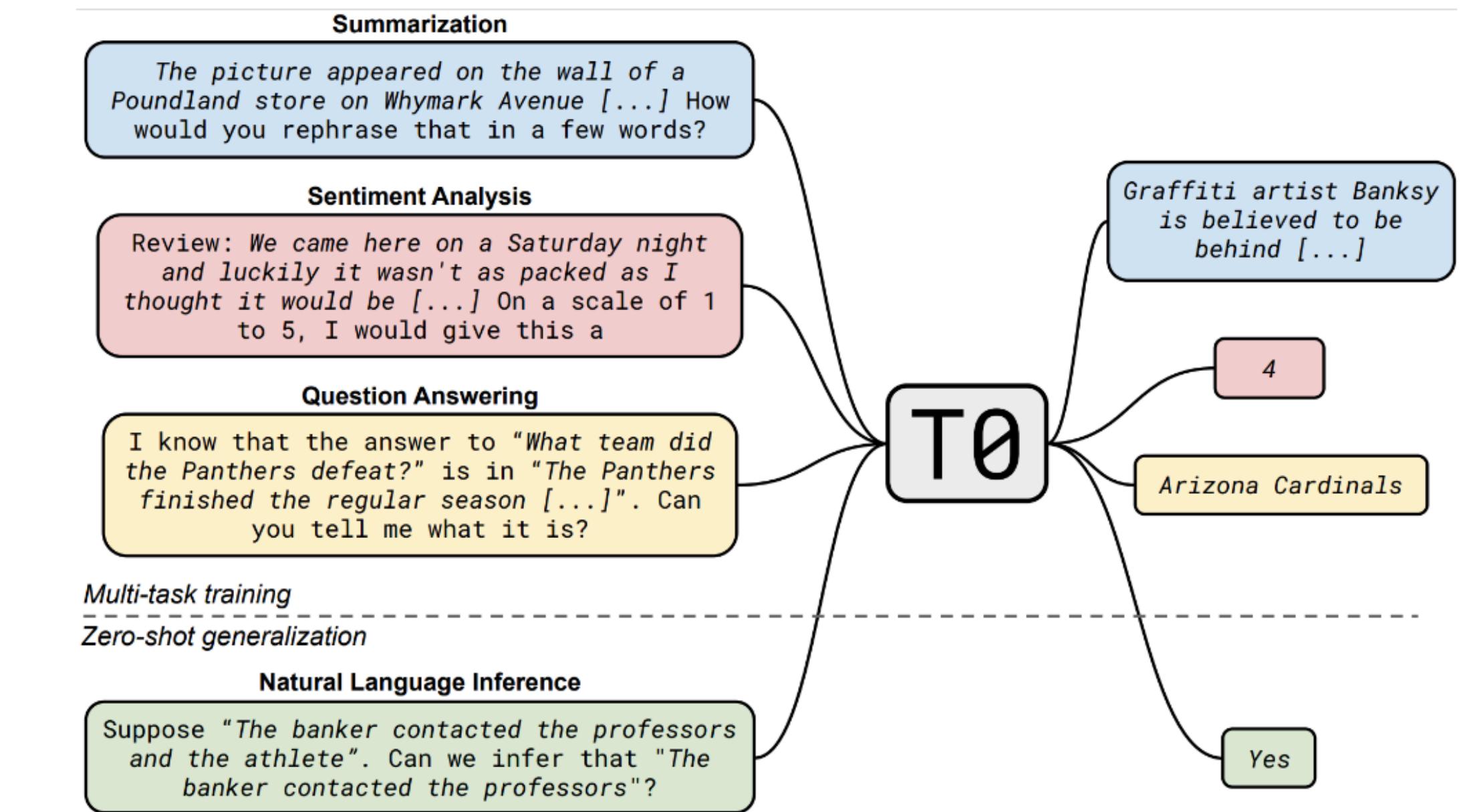
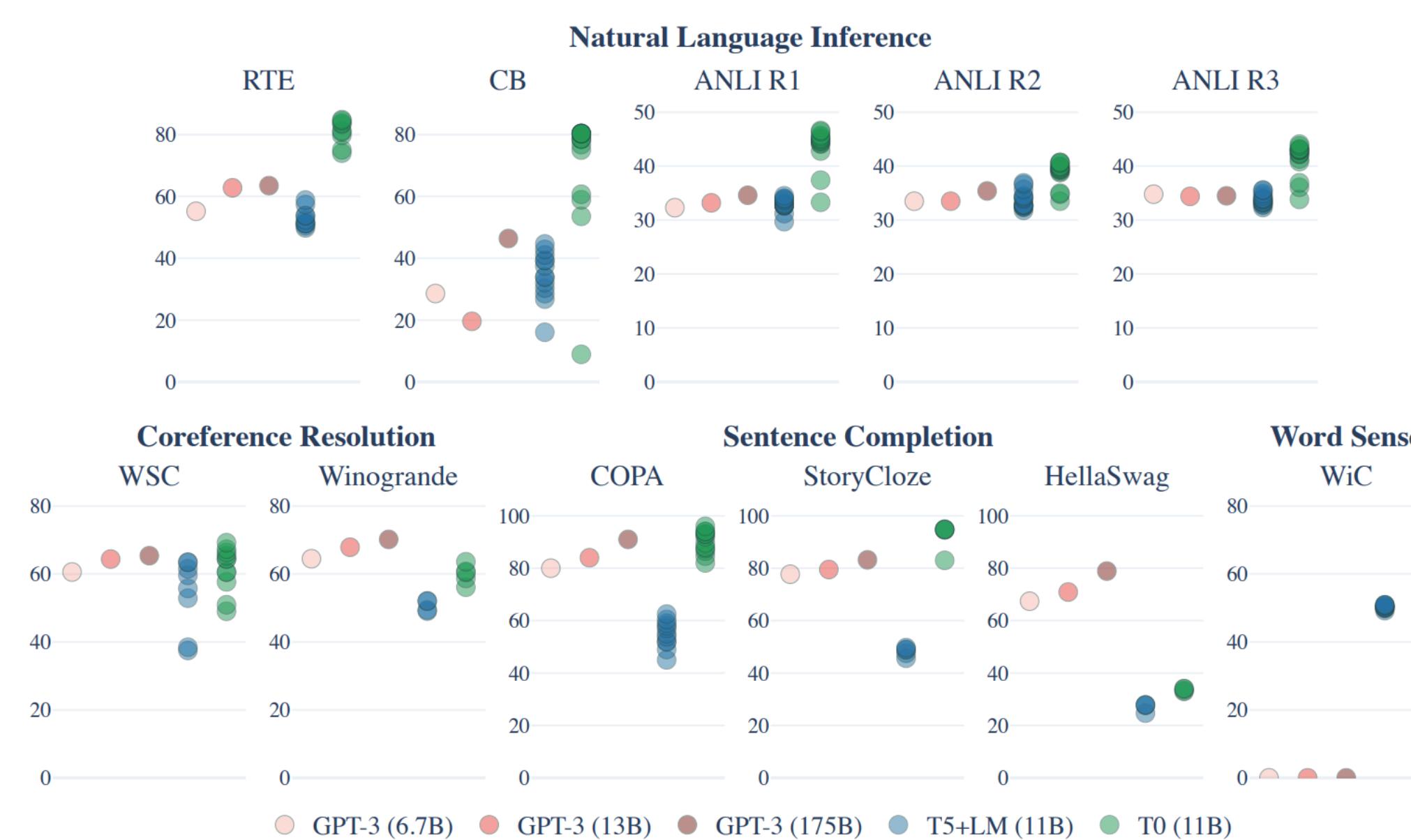


Figure 1. After training on a diverse mixture of tasks (top), T0 generalizes to unseen tasks (bottom).

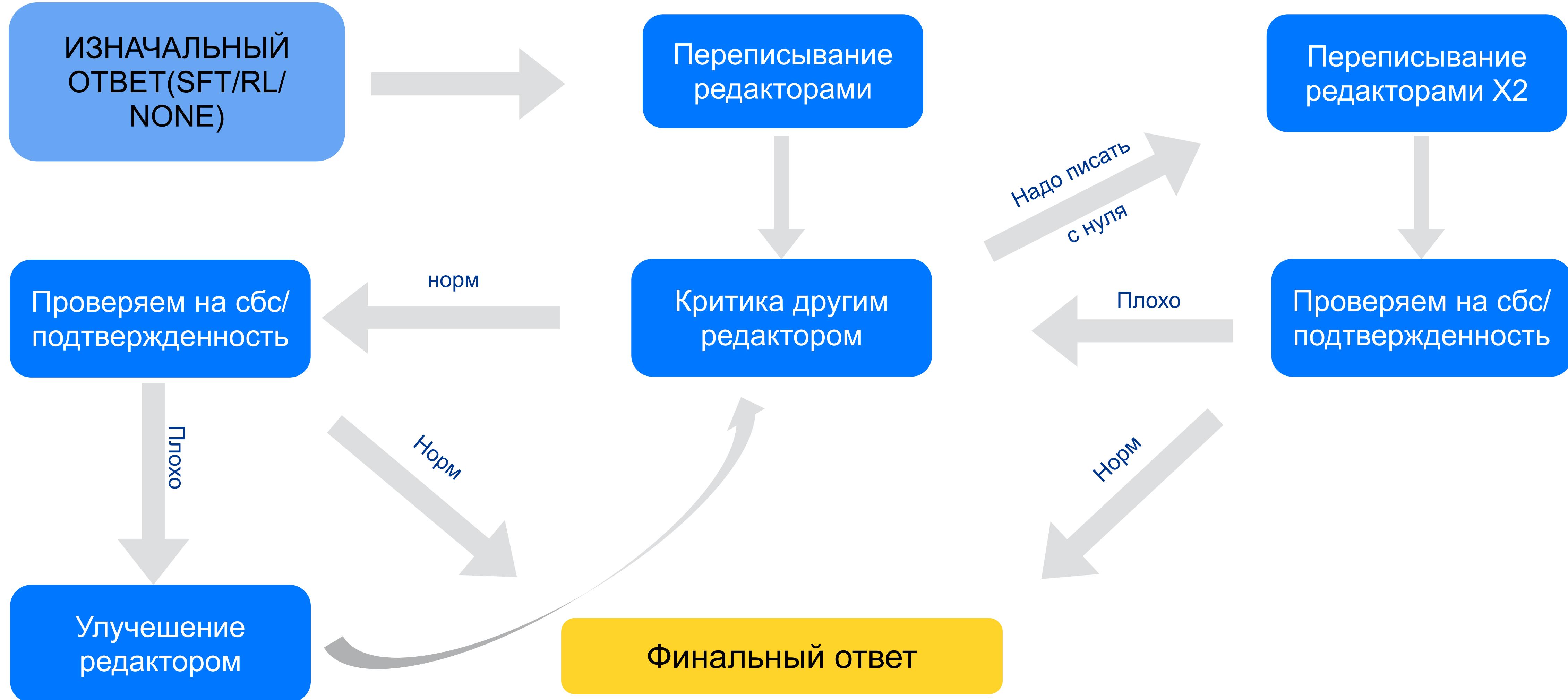
SFT

Как собирать инстракты

- Берете много запросов на которых хотите работать
- Сажаете много редакторов писать ответы на них
- ~~Profit~~

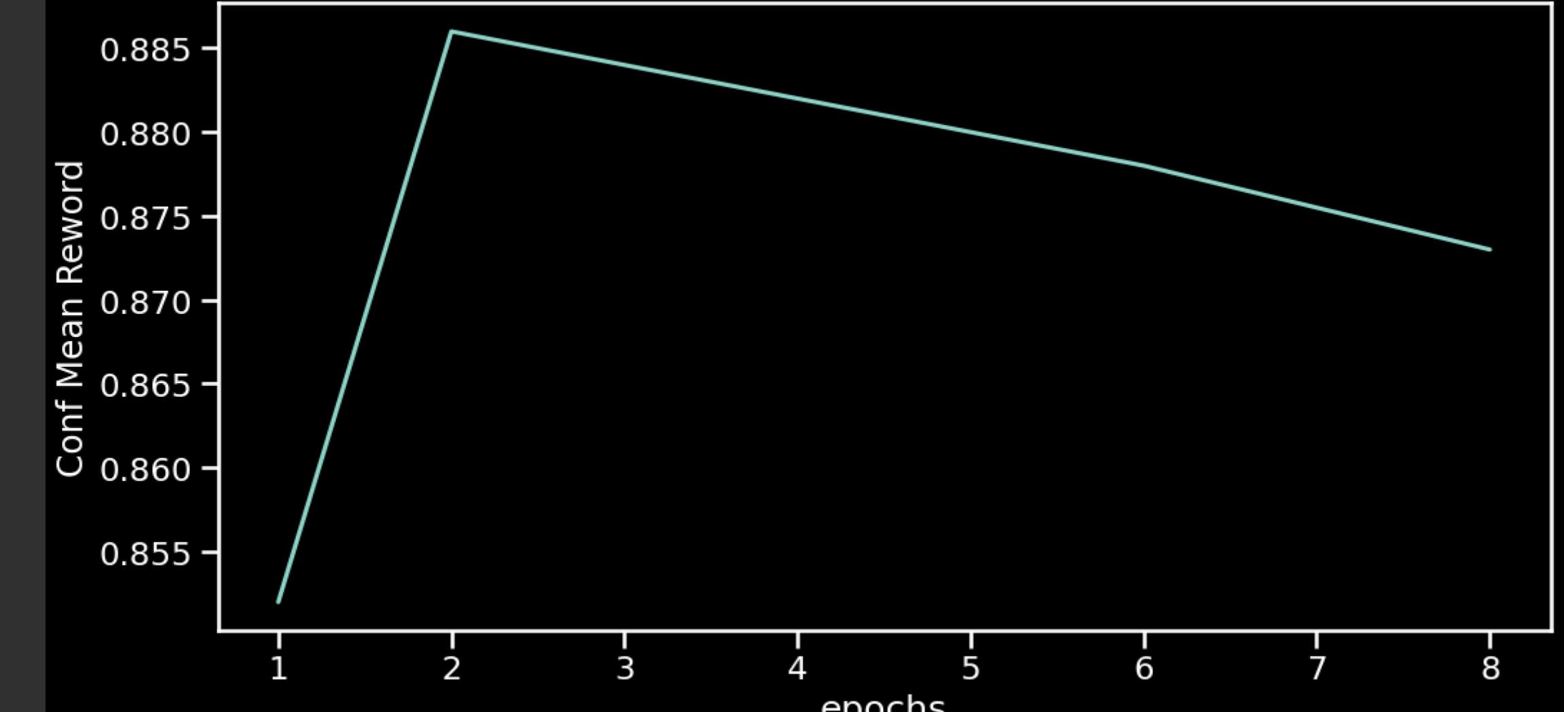
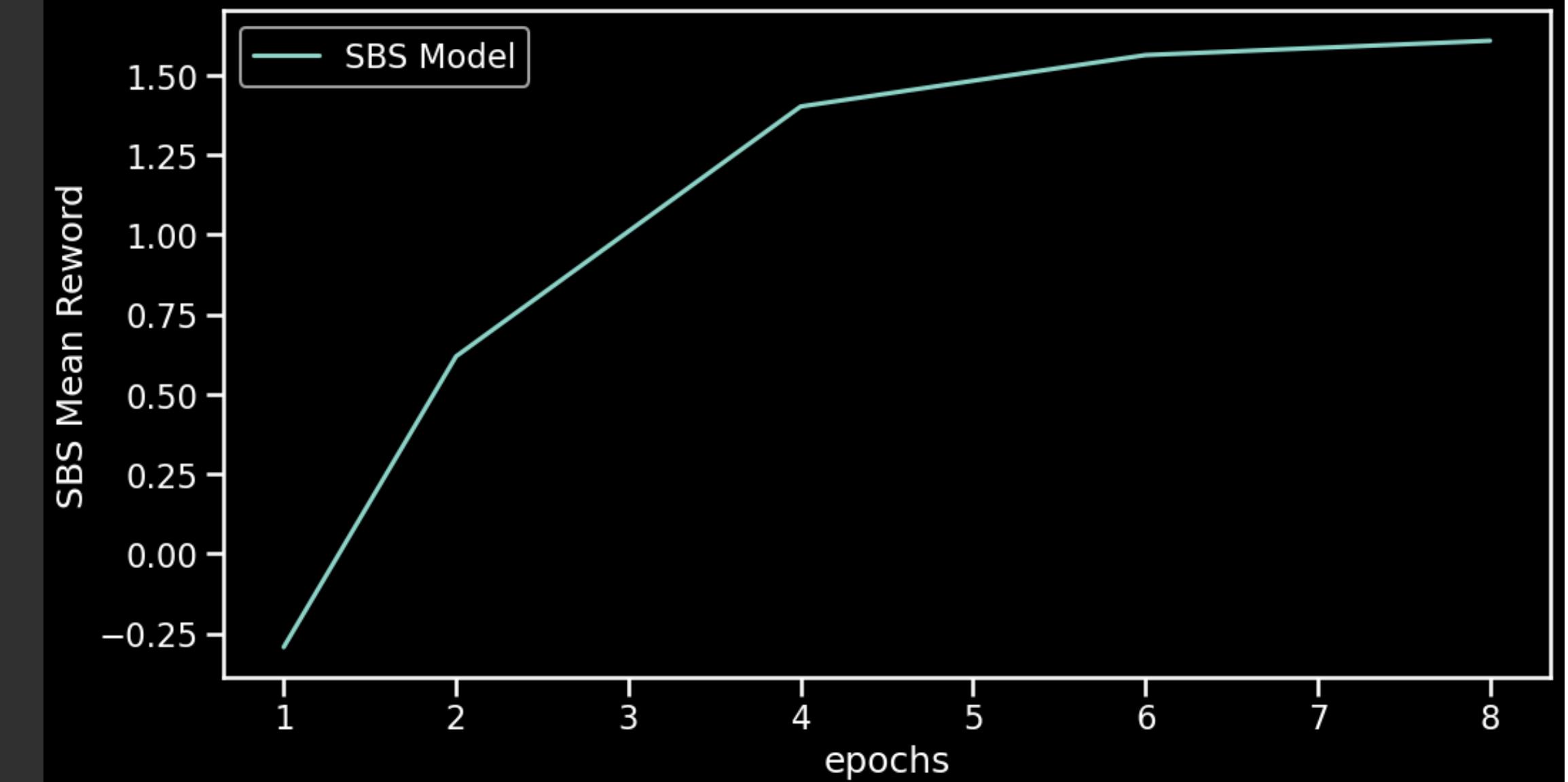
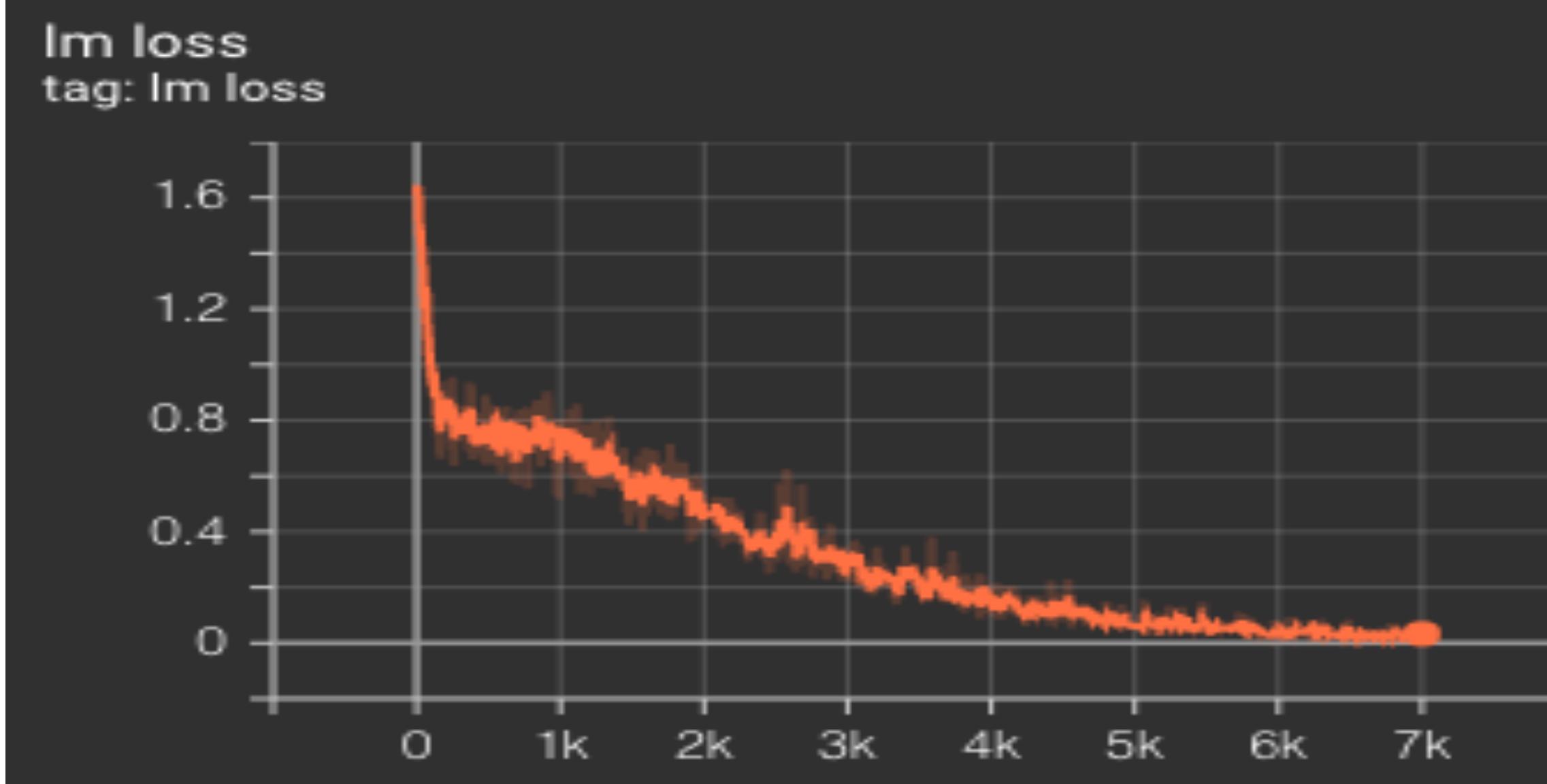
SFT

Как собирать инстракты



SFT

Как и сколько учить



SFT

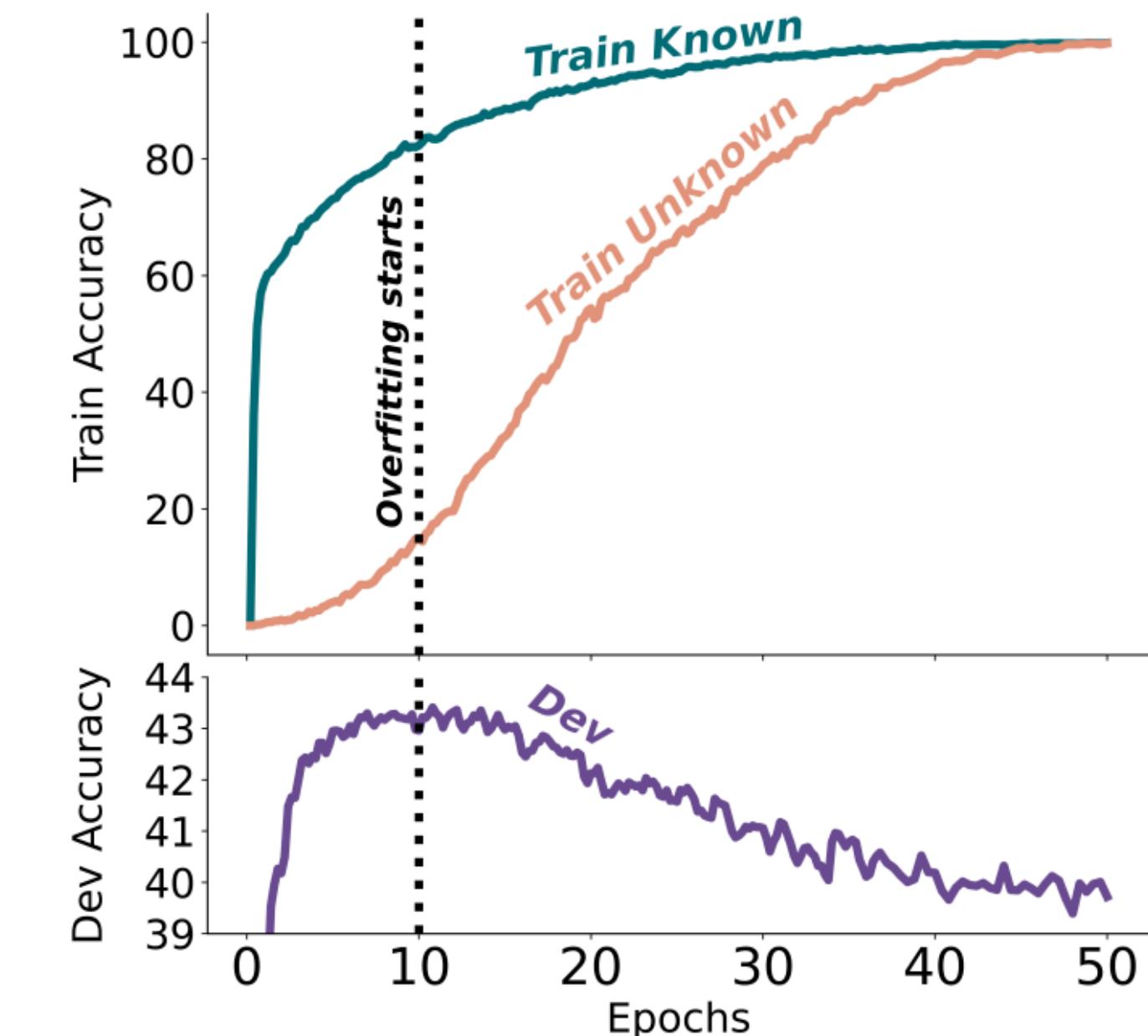
Does Fine-Tuning LLMs on New Knowledge Encourage Hallucinations?

Type	Category	Definition	Explanation
Known	HighlyKnown	$P_{\text{Correct}}(q, a; M, T = 0) = 1$	Greedy decoding <i>always</i> predicts the correct answer.
	MaybeKnown	$P_{\text{Correct}}(q, a; M, T = 0) \in (0, 1)$	Greedy decoding <i>sometimes</i> (but not always) predicts the correct answer.
	WeaklyKnown	$P_{\text{Correct}}(q, a; M, T = 0) = 0 \wedge P_{\text{Correct}}(q, a; M, T > 0) > 0$	Greedy decoding <i>never</i> predicts the correct answer, whereas temperature sampling with $T > 0$ <i>sometimes</i> predicts the correct answer.
Unknown	Unknown	$P_{\text{Correct}}(q, a; M, T \geq 0) = 0$	The model <i>never</i> predicts the correct answer, thus it seem to lack the knowledge of the correct answer.

(a)

Category	Question	Gold Answer	Greedy Answers	Sampled Answers
HighlyKnown	Who founded Science of Mind?	Ernest Holmes	[Ernest Holmes, .. Ernest Holmes, ..]	[..., ...]
MaybeKnown	What is the capital of Toledo District?	Punta Gorda	[Belmopan, .. Punta Gorda, ..]	[..., ...]
WeaklyKnown	What kind of work does Scott McGrew do?	Journalist	[Film director, .. Actor, ..]	[Musician, .. Journalist, ..]
Unknown	Where is Benedict located?	Hubbard County	[Louisiana, .. New Mexico, ..]	[Washington, .. Texas, ..]

(b)



SFT

Оверфит

model	% copypaste	From rewrite
batch1	31	None
batch2	38	None
batch5	42	June-rl
batch6	53	July-rl
batch7	60	Aug-rl
batch8	61	Sep-rl

RL

RL

Как проверить что заработает

- Сэмплинг может придумать ответ, лучше, чем beam search
- Реворд-модель этот ответ может обнаружить
- Аксессоры подтверждают, что ответ лучше, чем у beam search

	Confirmation	Usefulness
beam search	0.61	0.30
sampling t=0.7 top by rewards	0.67	0.37
sampling t=1.0 top by rewards	0.55	0.33

RL

Cross-Entropy RL/Rejection Sampling/Poor RL/Imitation Learning/...

- На каждый инстракт сэмплируем N траекторий
- Выбираем K лучших по reward модели
- Файнтюнимся

CERL

Плюсы и минусы

- **Идея:** если сэмплинг позволяет найти ответы лучше, чем у beam search - давайте на них обучимся

- **Алгоритм:**

- По пулу запросов сэмплировать набор ответов
- По каждому запросу выбрать лучший ответ по реворду
- Зафайнтюнится на выбранных ответах

- **Плюсы:**

- Можно проводить несколько итераций и растить качество
- Прост в реализации
- Хорошо работает как инит для других методов
- Можно использовать ответы других моделей
- Часто без него другие методы работают хуже

- **Минусы:**

- Неэффективен по сэмплам (для обучения используется 1 сэмпл из множества сгенеренных)
- Слабее других алгоритмов в долгом обучении



- Ox - итерация cerl, Oy - реворд
- Зеленая - среднее максимальное значение реворда по сэмплам
- Синяя - значение реворда на эвале (бимсерч)
- Красная - значение реворда на первне (бимсерч)

RL

Формула

- Есть много аспектов качества: полезность, подверженность, компетентность, безопасность, орфография и тд
- Есть продуктовые требования - длины ответа 150-160 токенов, поменьше одинаковых блоков ссылок, минимизировать кол-во ответов “Я не знаю” и тд
- Хочется оптимизировать их все
- Есть main аспект в sbs но его не достаточно

RL

Формула

- Reword_part =
sbs_main * a1
sbs_usefulness * a2
sbs_grades_confirmation * a3
confitmation * a4
- regexp_part =
 $F(\text{few_footnotes}) + F(\text{tokens_count}) + F(\text{num_brackets}) + F(\text{is_getoff})$
 F - по началу индикаторная, сейчас сглаживающая индикатор функция
- Есть умные статьи которые не работают, например <https://arxiv.org/pdf/2402.00742>

RL

Вводная

- $\pi_\theta(a | s)$ - модель которая выдает ответ a по входу s
- $R(a, s)$ - реворд для ответа a по входу s
- $V(s) = E_{x \in \pi, s} R(x)$ - матожидание набранной награды из состояния s
- $Q(a, s) = r(a, s) + \gamma \cdot E_{s'} V(s')$ - матожидание набранной награды из состояния s , совершая действие a
- $A(a, s) = Q(a, s) - V(s)$

RL

Policy Gradient

- $J(\pi_\theta) = E_s E_{a \in \pi_\theta} R(a, s)$
- <Пропускаем формулы про вывод Policy Gradient>
- $\nabla_\theta J(\pi_\theta) = E_s E_{a \in \pi_\theta} \nabla_\theta \log \pi_\theta(a | s) R(a, s)$

RL A2C

- Заводим отдельную V голову/модель
- $\nabla_{\theta} J(\pi_{\theta}) = E_s E_{a \in \pi_{\theta}} \nabla_{\theta} \log \pi_{\theta}(a | s) [R(a, s) - V(s)]$
- V модель учим на $E_s E_{a \in \pi_{\theta}} (V(s) - R(a, s))^2$

<https://arxiv.org/abs/2004.14288>

RL A2C

- Вход: множество запросов для обучения D
- Инициализируем политику SFT-моделью: $\pi_\theta \leftarrow \pi_{SFT}$
- Инициализируем ценность и модель награды: $V_\phi \leftarrow R$
- Повторять до сходимости
 - Выбираем батч запросов $B \in D$
 - Вычисляем ценность для каждого запроса из батча V_ϕ
 - Генерируем по одному ответу a_i на каждый запрос. Важно генерировать именно актуальной обучаемой моделью π_θ
 - Вычисляем награду $R(s_i, a_i)$ для всех пар
 - Вычисляем лосс для агента $L_a = -\frac{1}{B} \sum_i^B \log \pi_\theta(a_i | s_i) [R(s_i, a_i) - V_\phi(s_i)]$
 - Вычисляем лосс для функции ценности V - $L_v = \frac{1}{B} \sum_i^B [V_\phi(s_i) - R(s_i, a_i)]^2$
 - $(L_a + L_v).backward()$

RL PPO

- Вспомним про GAE
- Будем стандартизировать advantage по бачку
- Будем делать importance sampling
- Будем клипать градиенты чтобы не разваливаться
- Последний год не видел статей в которых используют PPO...

<https://arxiv.org/pdf/1707.06347>

RL DPO

- Допустим что $E_a[r_\phi(s, a) - \beta KL(\pi_\theta(a | s), \pi_{SFT}(a | s))] - > \max_\theta$
- Тогда мы можем выписать оптимальную политику
 - $\pi^*(a | s) = \frac{1}{Z(s)} \pi_{SFT}(a | s) e^{\frac{1}{\beta} r(s, a)}, Z(s) = \sum_a e^{\frac{1}{\beta} r(s, a)}$
 - Выразим реворд $r(s, a) = \beta \log\left(\frac{\pi^*(a | s)}{\pi_{SFT}(a | s)}\right) + \beta \log Z(s)$

RL DPO

- $r_\theta(s, a) = \beta \log\left(\frac{\pi_\theta(a | s)}{\pi_{SFT}(a | s)}\right) + \beta \log Z(s)$. функция награды в которой π_θ оптимальна
- $r_\theta(s, a) = \beta \log\left(\frac{\pi_\theta(a | s)}{\pi_{SFT}(a | s)}\right)$
- $\sum_{(s, winner, loser) \in D} \log \sigma(r_\theta(s, winner) - r_\phi(s, loser)) - > \max_\theta$ - Бредли Терри
- $\sum_{(s, winner, loser) \in D} \log \sigma(\beta[\log\frac{\pi_\theta(winner | s)}{\pi_{SFT}(winner | s)} - \log\frac{\pi_\theta(loser | s)}{\pi_{SFT}(loser | s)}]) - > \max_\theta$

RL

DPO на синтетике

- Генерируем 50 сэмплов на промыт
- Берем лучший vs худший
- Модель улучшается но слабо
- Можно ли лучше ?

RL

DPO на синтетике

- Генерируем 50 сэмплов на промпт $a_1, a_2 \dots a_{50}$
 - Берем 2 случайных a_1, a_2
 - $r_1 = R(a_1, s); r_2 = R(a_2, s)$
 - $r_{gap} = \sigma(r_1 - r_2)$
 - Если $r_{gap} > threshold$ то добавляем в обучающий пул

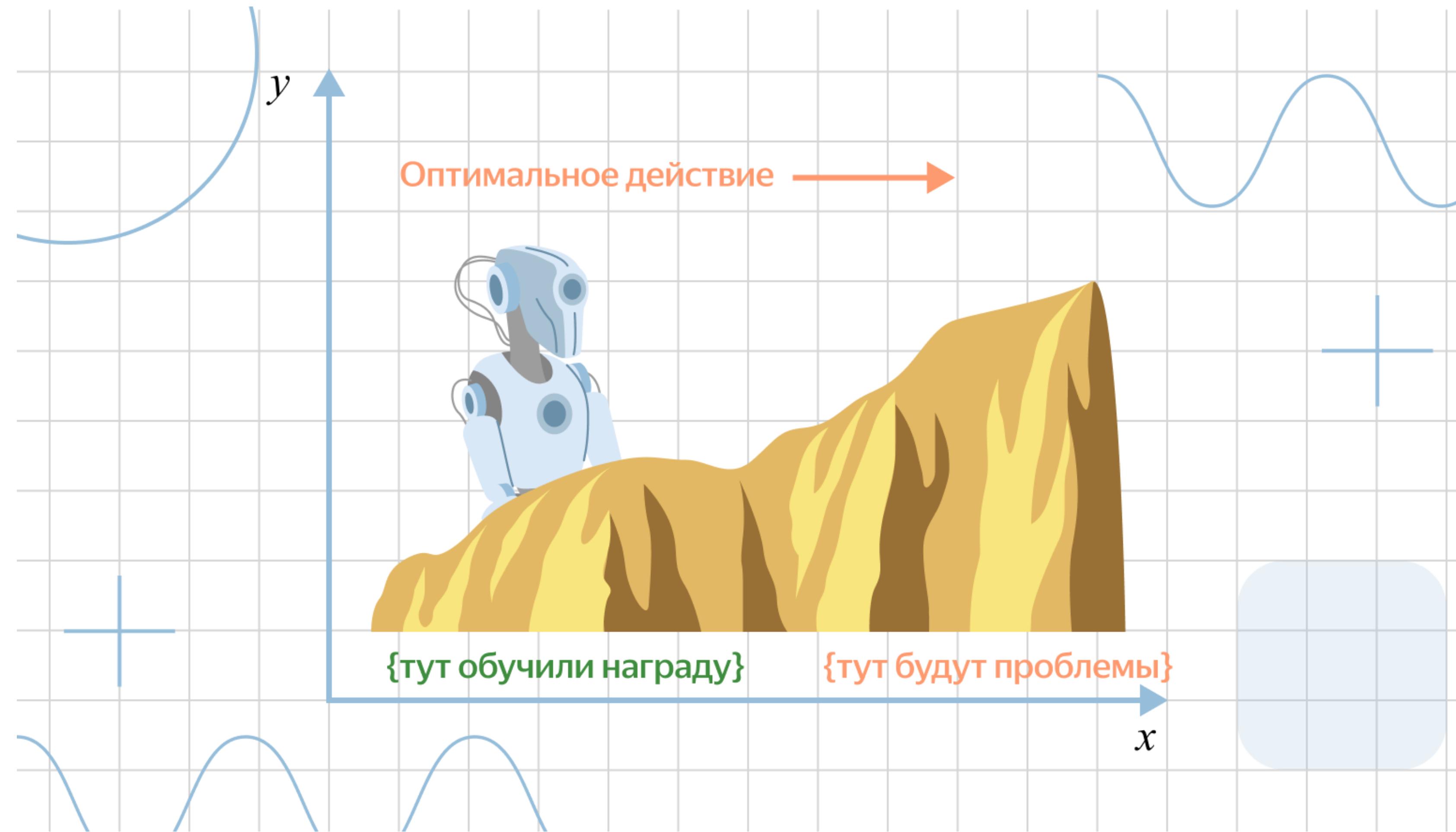
RL

DPO, вариации

- DPO: $\sum_{(s, winner, loser) \in D} \log \sigma(\beta [\log \frac{\pi_\theta(winner | s)}{\pi_{SFT}(winner | s)} - \log \frac{\pi_\theta(loser | s)}{\pi_{SFT}(loser | s)}]) \rightarrow \max_\theta$
- CPO: DPO + NLL_loss
- Часто меняют $\pi_{SFT} \rightarrow EMA(\pi_{SFT}, \pi_\theta)$

Проблемы RL

Доменний сдвиг



Проблемы RL

Доменний сдвиг, академический подход

- Добавляем $KL(\pi_\theta(a | s), \pi_{SFT}(a, s))$ в лосс RL и далеко не узжаем
- На самом деле почти у всех лучше работает добавлять вместо π_{SFT} - $EMA(\pi_\theta, \pi_{SFT})$

Проблемы RL

Доменный сдвиг, “нормальный” подход

- KL штраф лишь замалчивает проблему вашего реворда
- Нужен бесконечный active learning
 - Обучили RL
 - Оверфитнули реворд, поняли это по разметкам
 - Залили в разметки кучу генераций нового rl против sft/старой модели/стадии rl перед началом оверфита. Важно заливать не только гриди но и сэмплинги/эмулирование пика лучшего
 - Переобучили реворд модели
 - Повторяем все заново
- У нас сейчас 37-я версия реворд модели, в каждой версии исправлялись какие-то баги и неточности прошлых версий

Проблемы RL

Качество реворд модели, что размечать

Какие батчи хорошие:

- **Сложные** для текущих ревордов: плохой accuracy, skew
- **Полезные**: Ответы, которые мы скорее хотим показывать

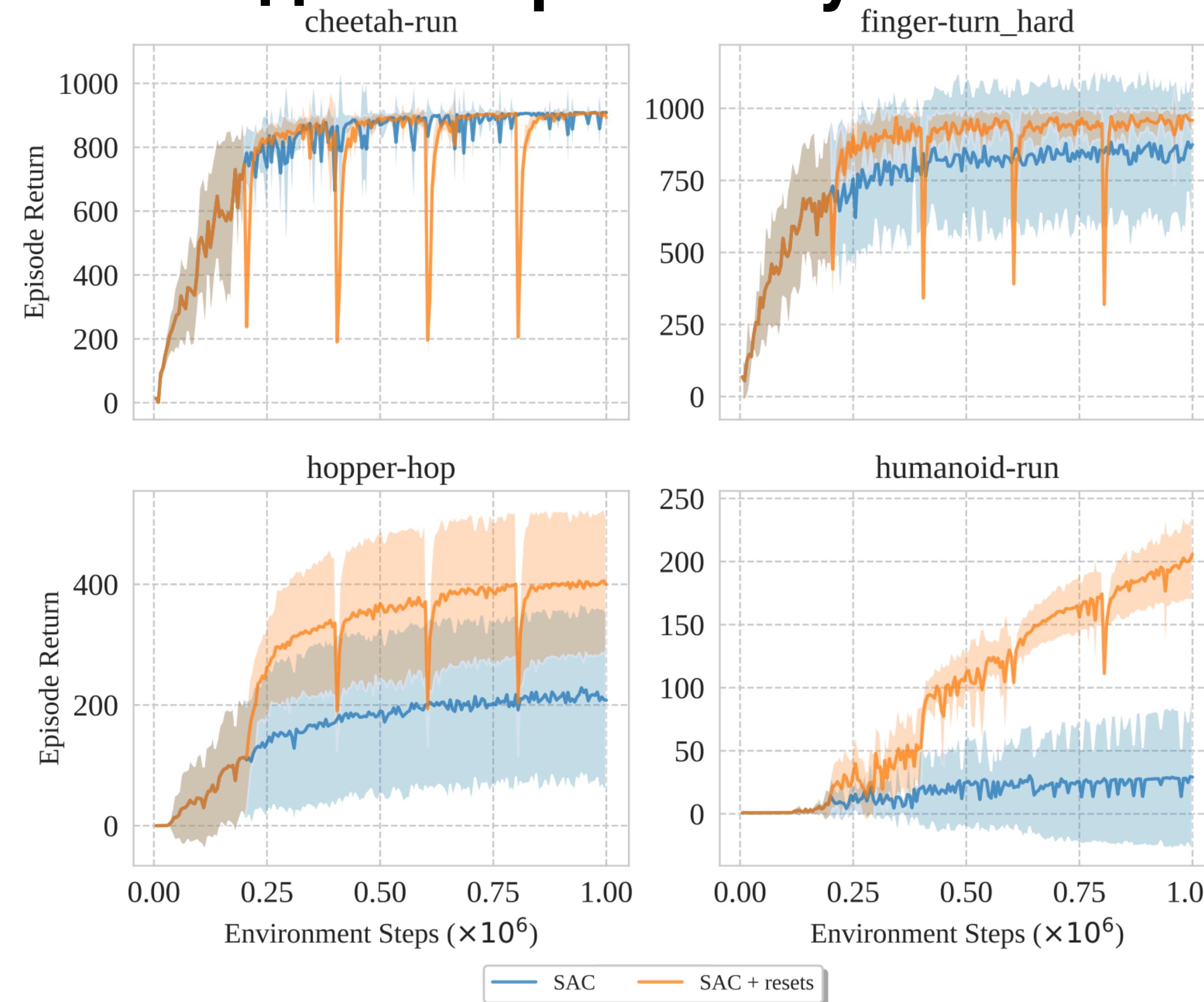
Что перебираем в разметках:

- **Модели**: SFT, RL-ные модели всех стадий, ответы написанные людьми
- **Способы сэмплинга**: бимсёрч, случайный сэмплинг, топ-1 по реворду/формуле
- **документы**: одинаковые/разные, топы по ревордам, разные модели ранкера и формулы ранжирования
- **Запросы**: есть сложные запросы, где весь стек моделей не справляется

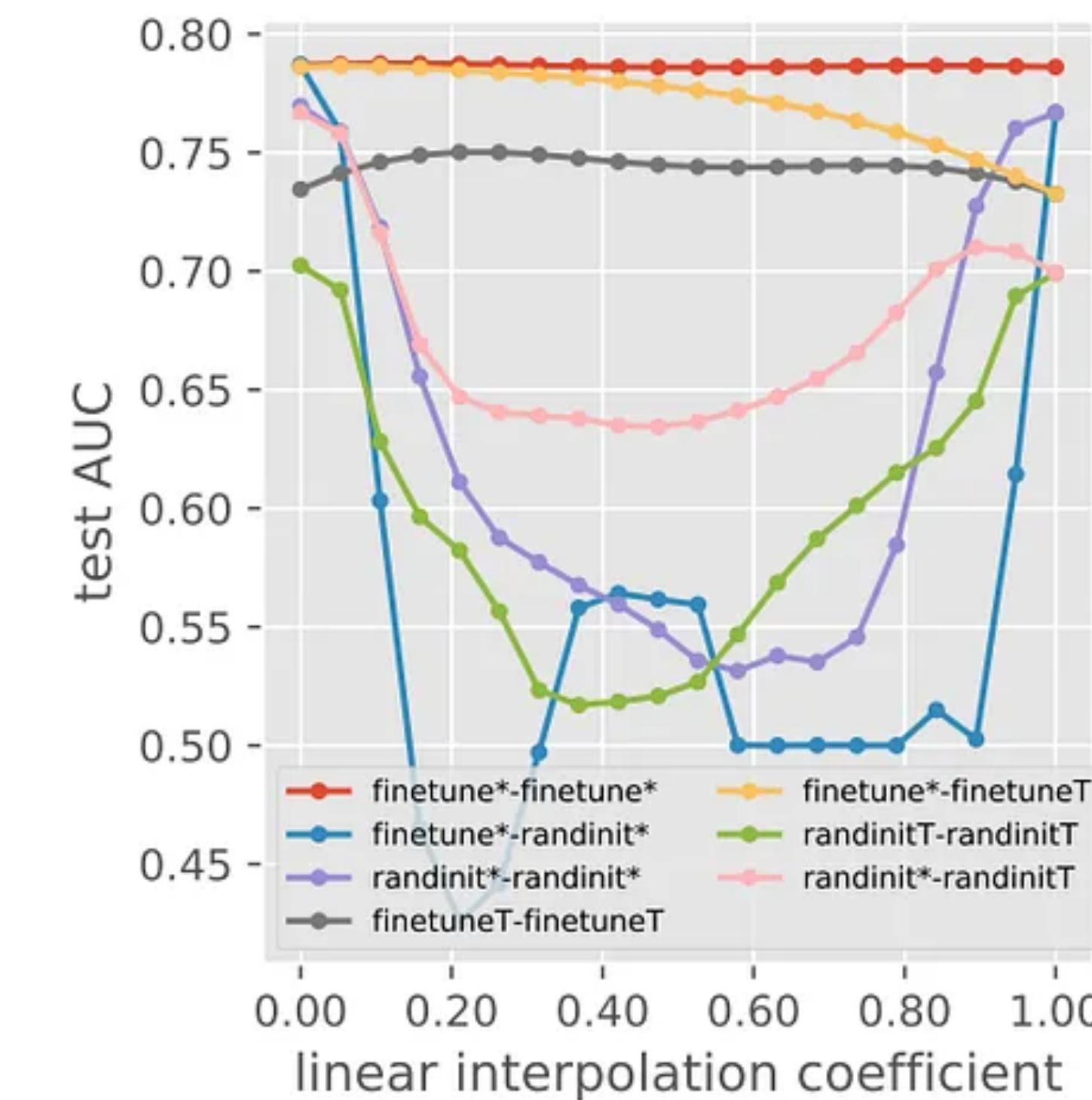
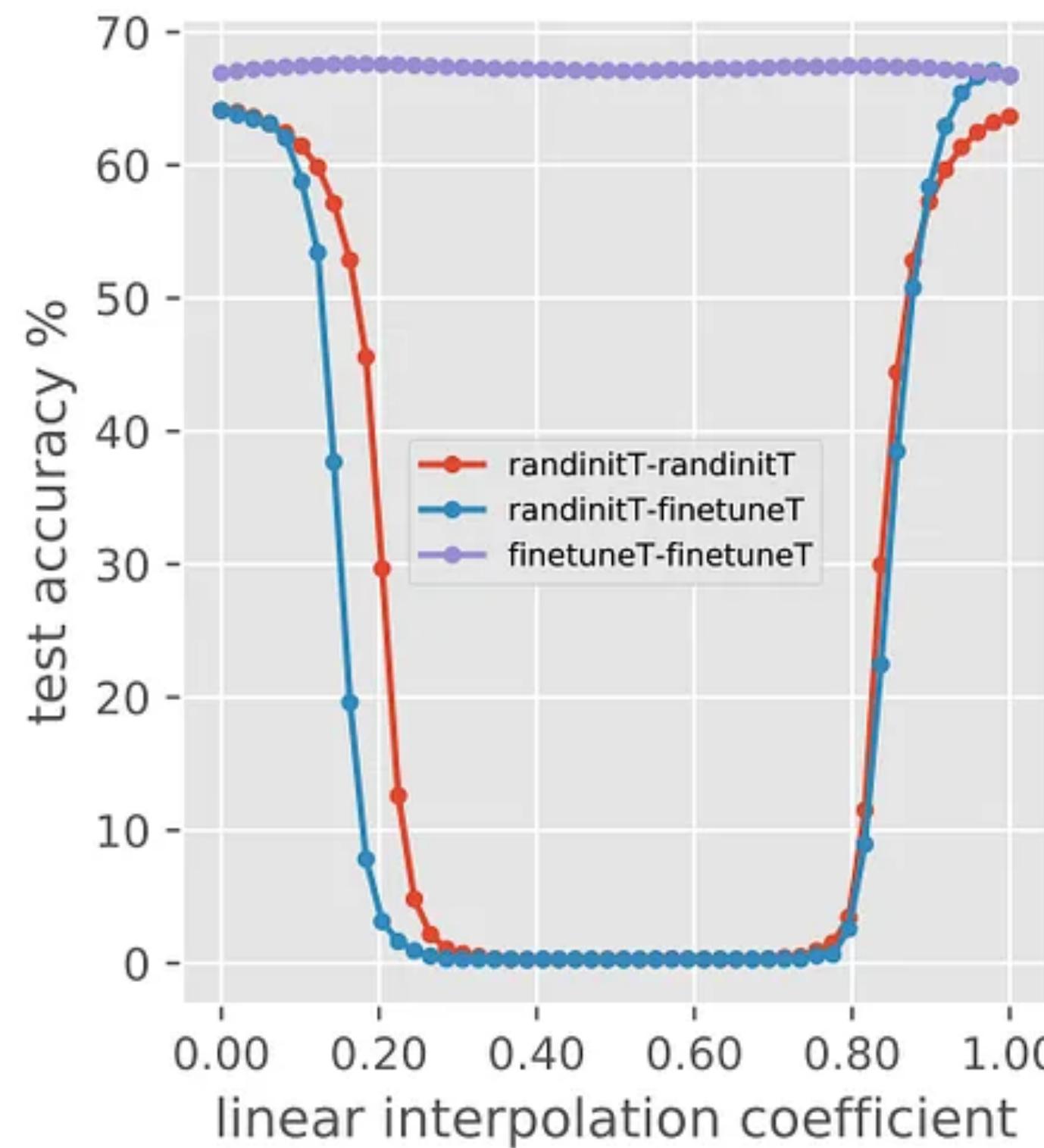
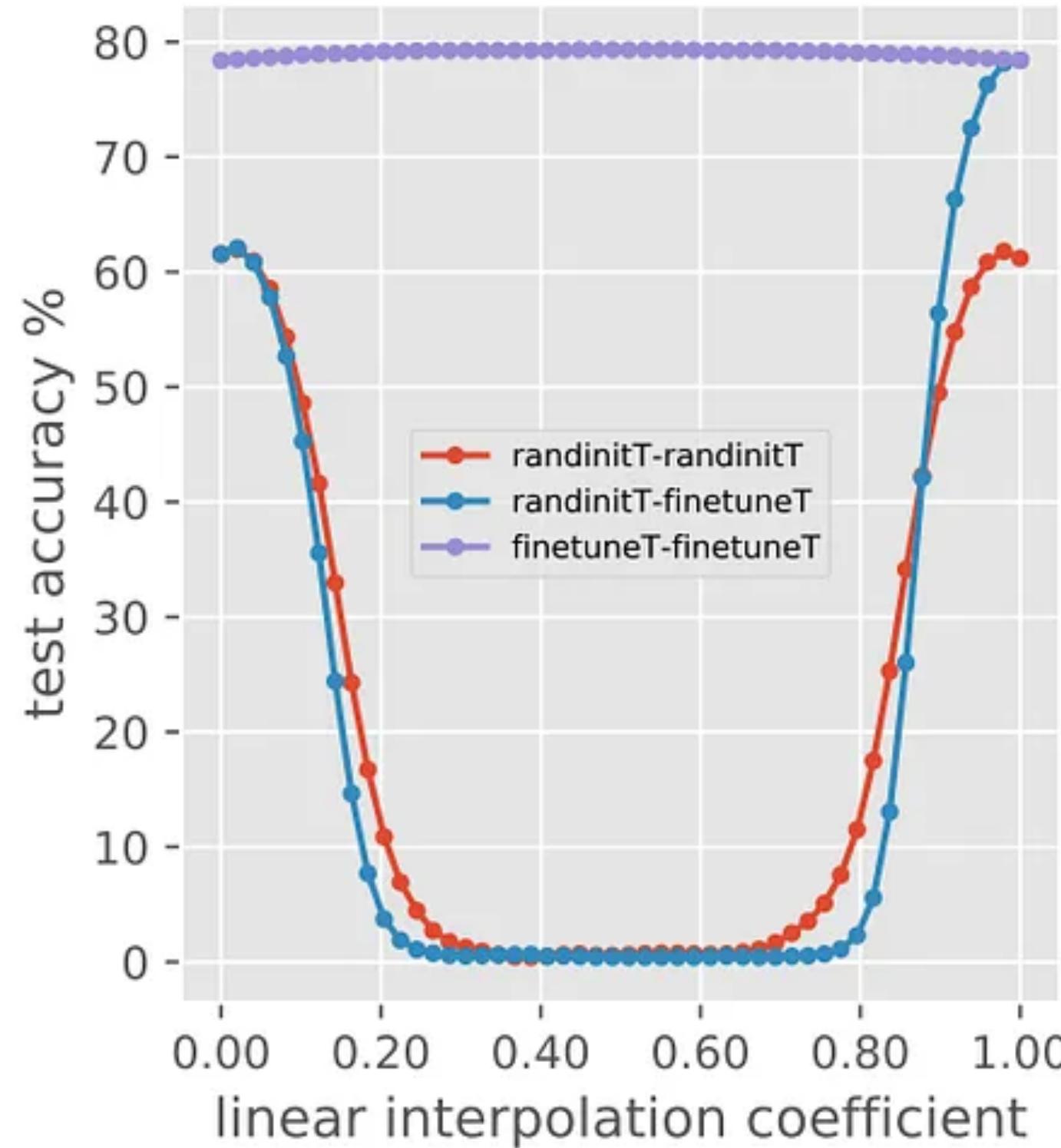
Всё это x2! Очень много вариантов!!!

Проблемы RL

Модель перестает учиться



General Linear mode connectivity



General TIES

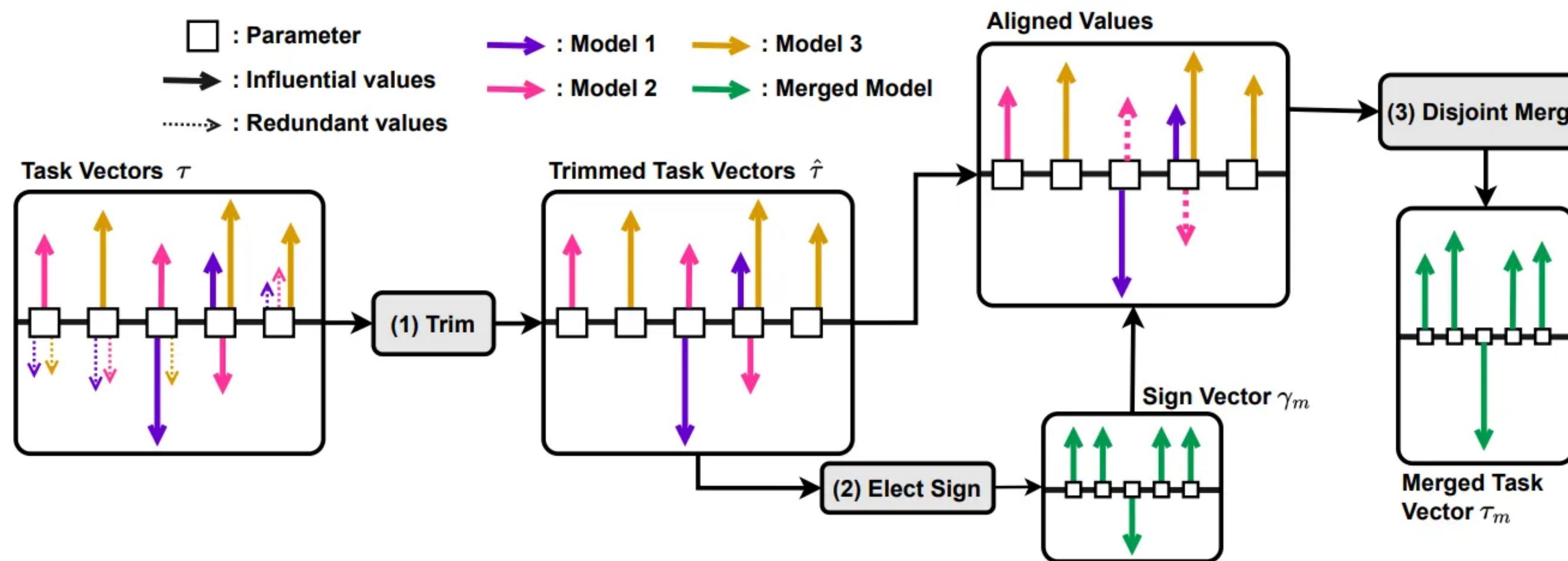
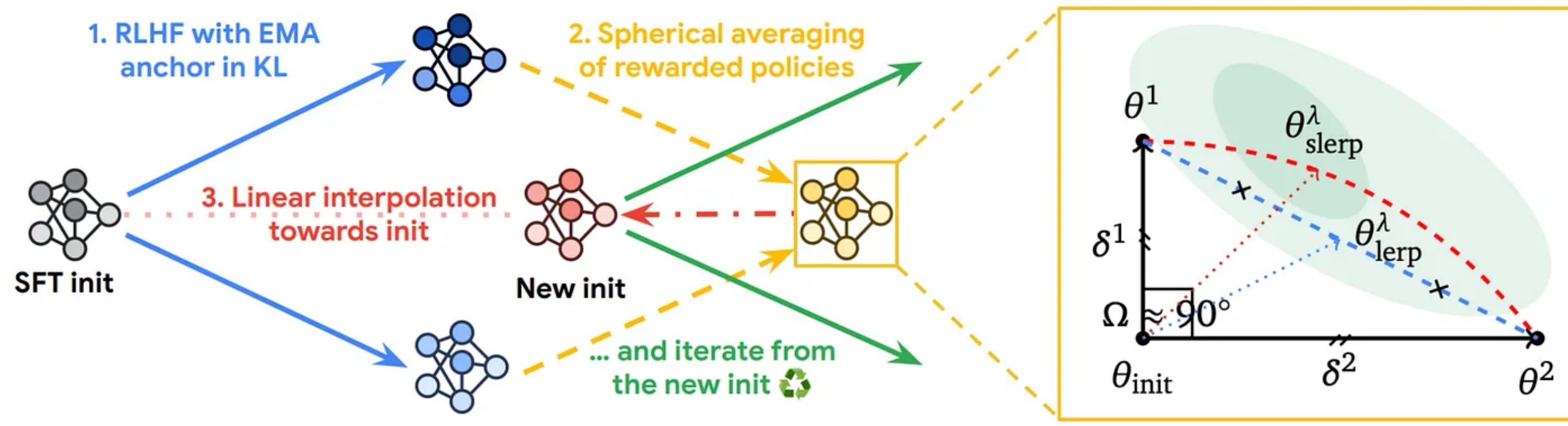


Figure 1: A depiction of the steps involved in TIES-MERGING. We visualize each parameter in a model as a square. The arrows depict the update (task vector, τ) to a parameter produced by fine-tuning on different tasks (coded by colors), with direction denoting sign and length denoting magnitude. We first *trim* the task vector values based on their magnitude, then we *elect* the sign for each parameter (γ_m , green vector containing +1 or -1) by resolving sign conflicts. Finally, we pick only the values that align with the elected sign and take their mean as the final parameter value.

<https://arxiv.org/abs/2306.01708>

Подход к RL в GEMMA-2

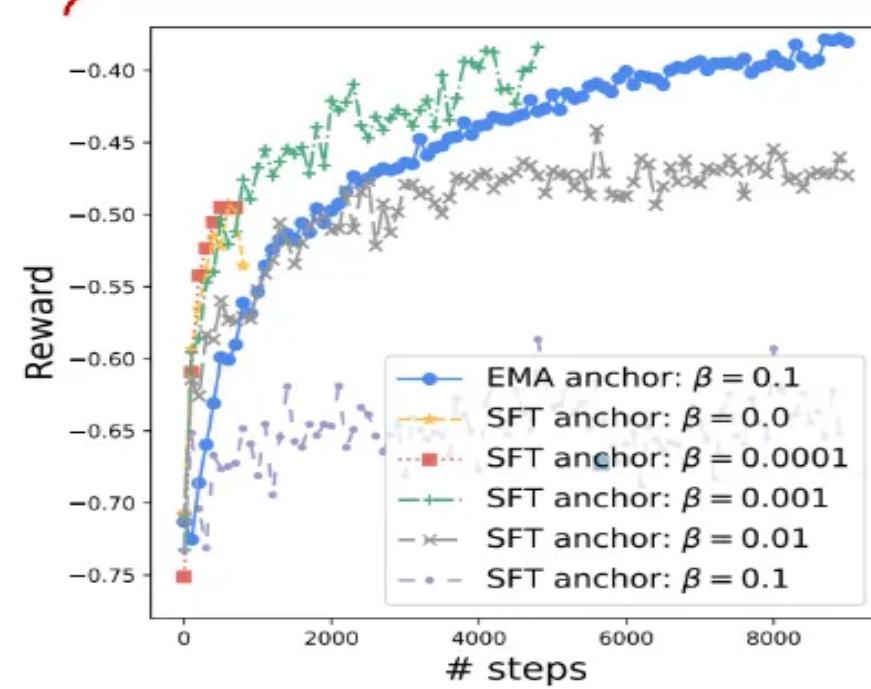
WARM



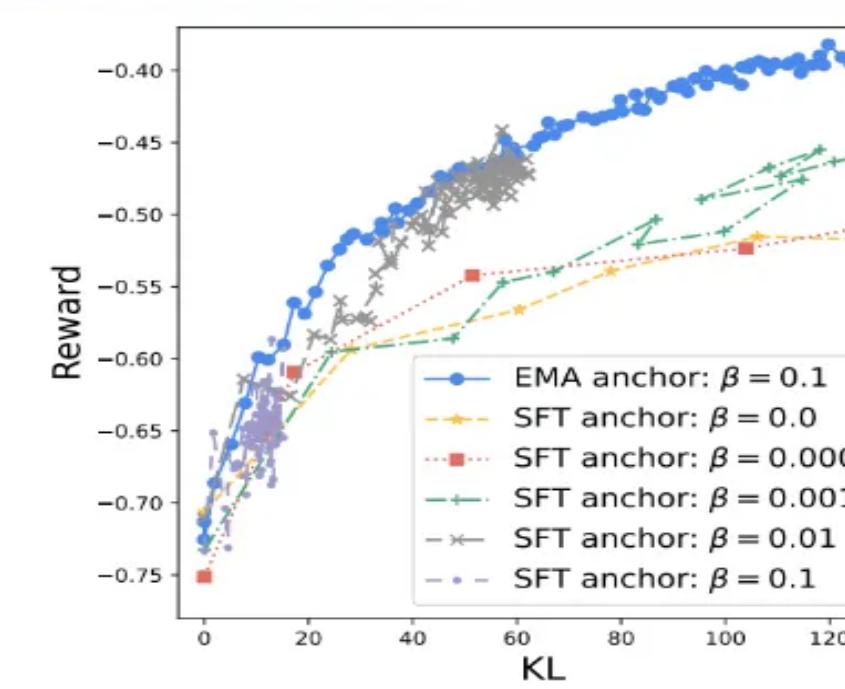
WARP with three model merging stages, applicable iteratively.

Подход к RL в GEMMA-2

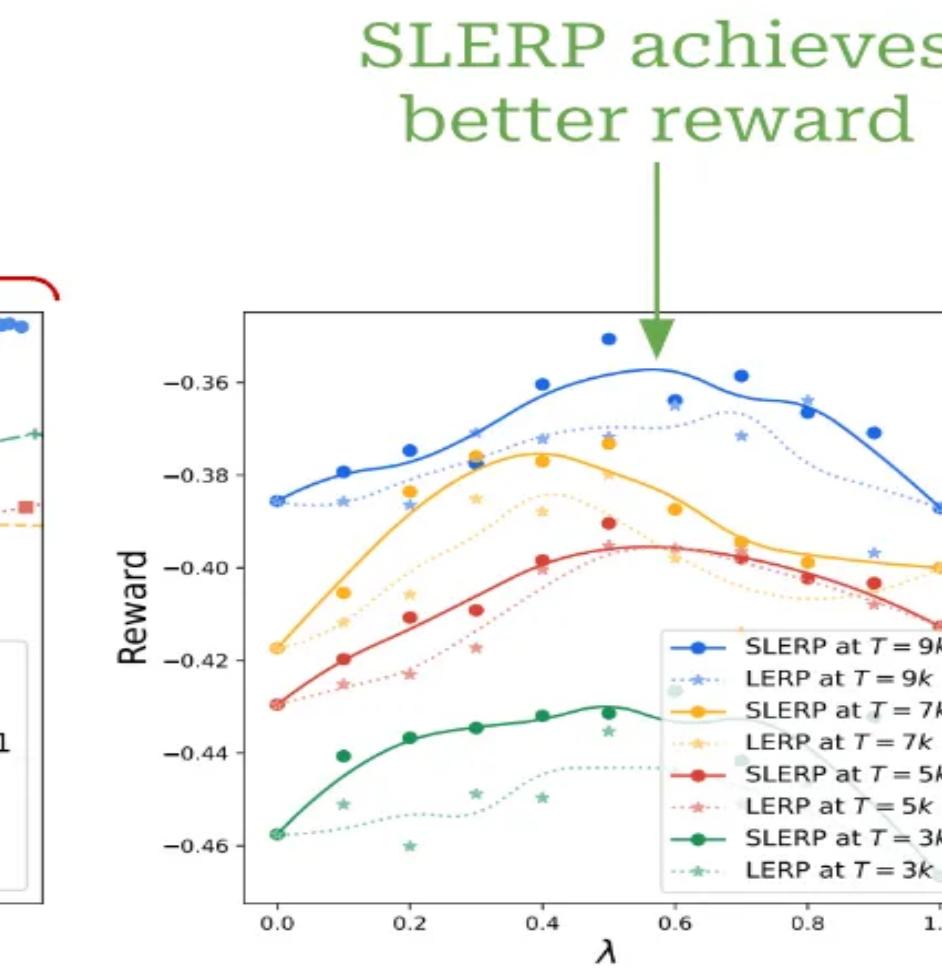
Using an EMA anchor yields a better tradeoff between reward / KL



(a) Reward vs. steps.

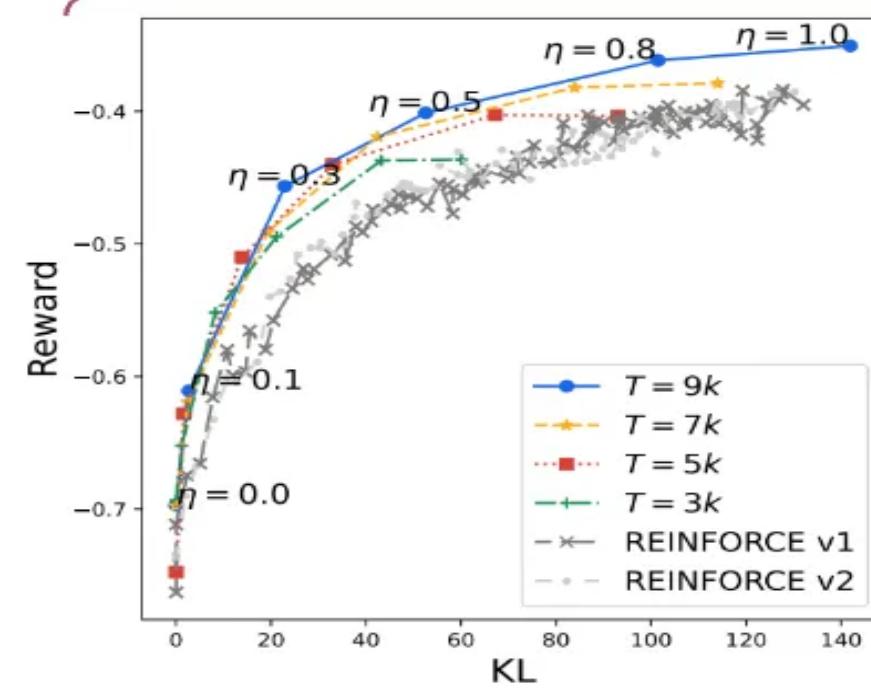


(b) Reward vs. KL.

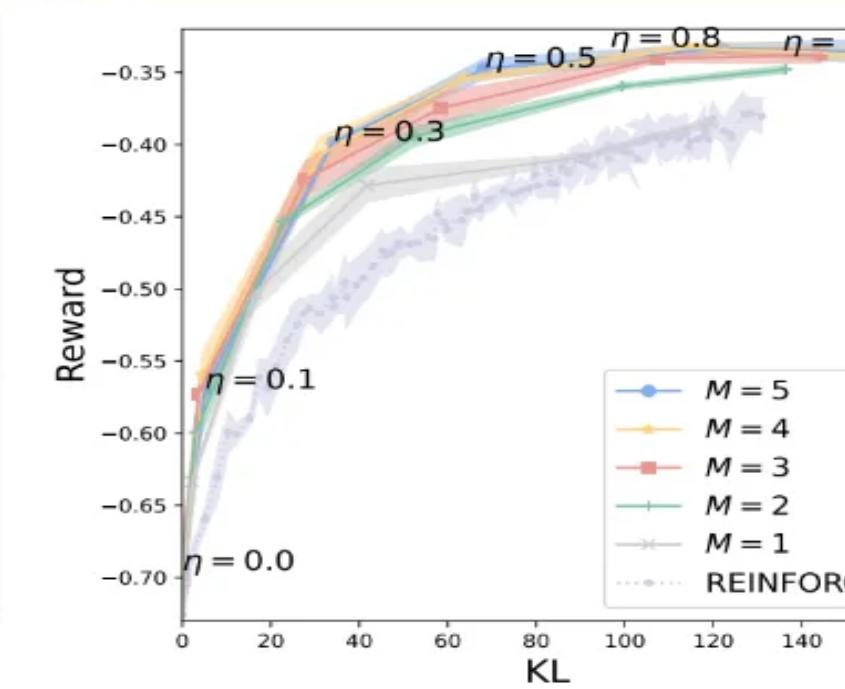


(c) SLERP vs. LERP.

Merging multiple independently trained policies and interpolating towards the SFT initialization is helpful!

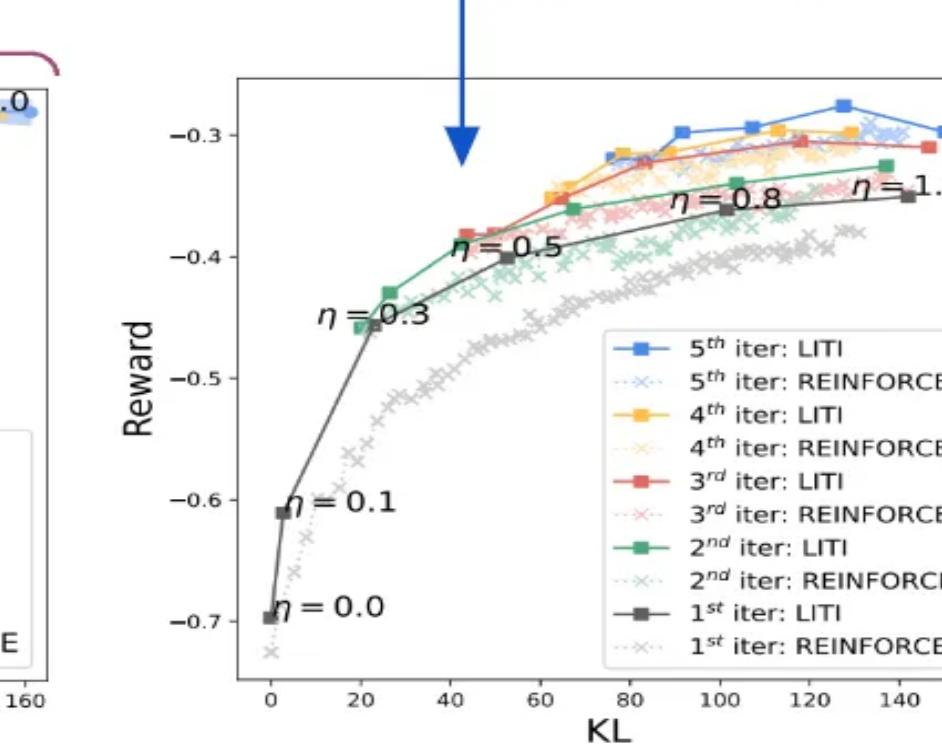


(a) LITI of SLERP after T steps.



(b) LITI of SLERP of M weights.

WARP should be applied iteratively

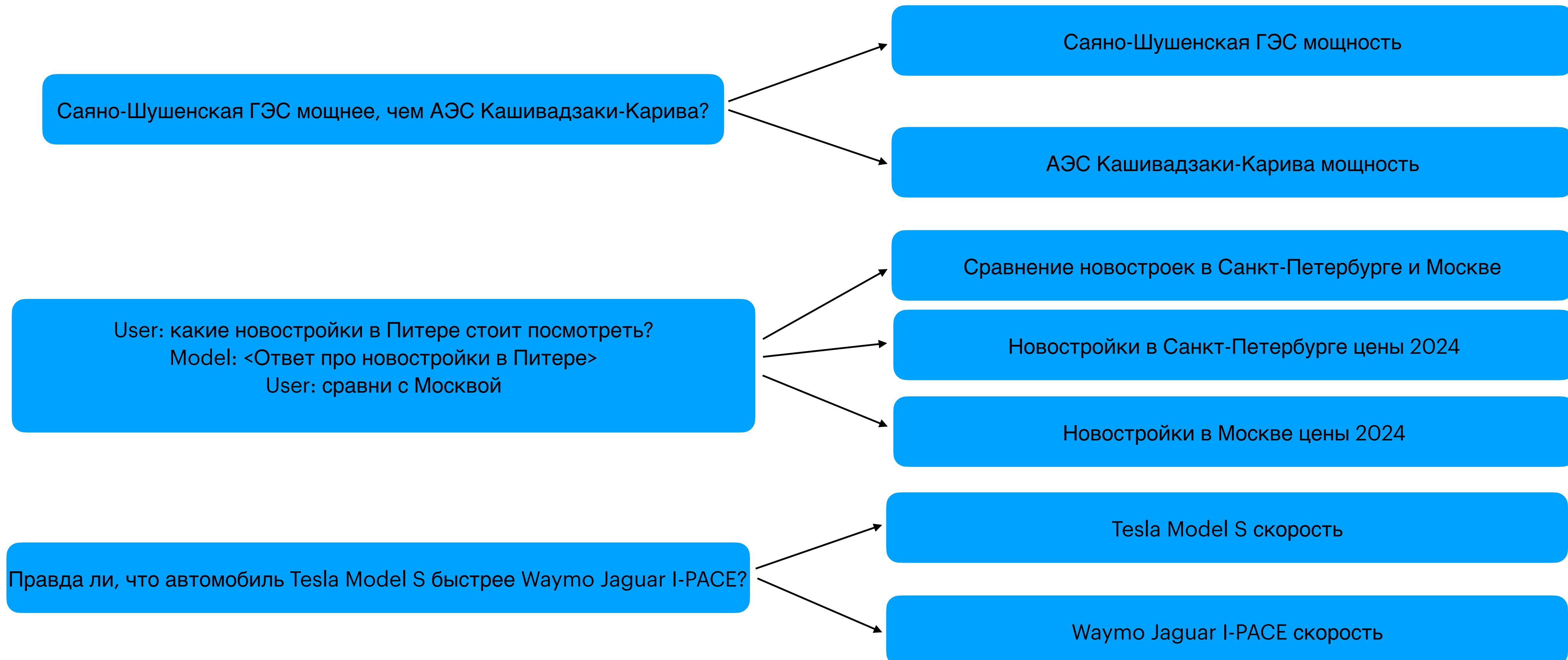


(c) Iterative WARP.

RAG

Рефразер

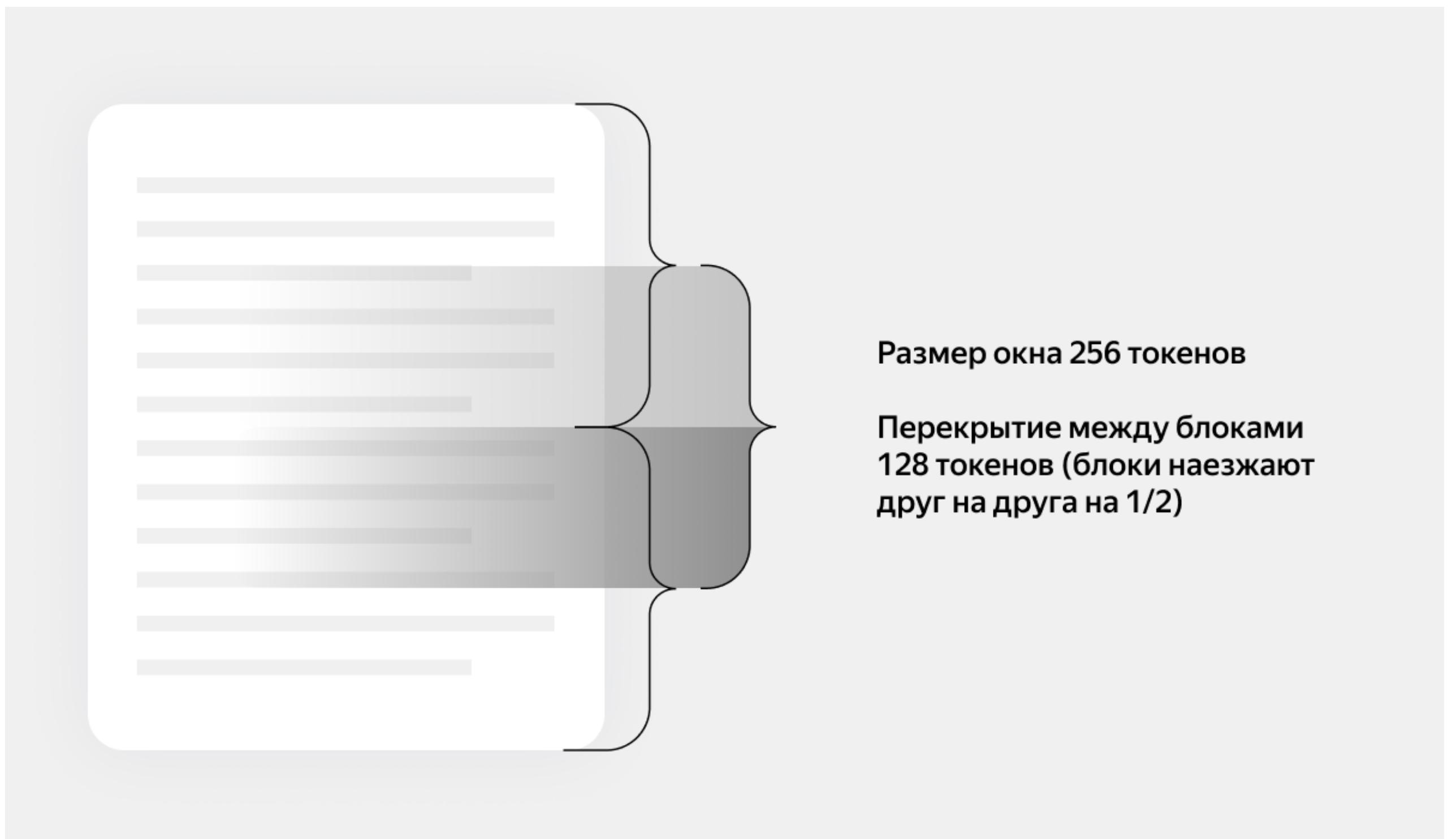
Как найти нужные документы



Документ

InfoContext

- Текст документа слишком огромный
- Разобьём его на блоки длиной 256 токенов с перекрытием 128
- Поскорим блоки по релевантности запросу с помощью очень маленькой BERT-модели (8 слоёв).
- Выбираем самые релевантные блоки, пока не наберём 1500 токенов уникального текста.
- «Склейываем» выделенные блоки в том же порядке, в котором они появились в исходном тексте.



Reranker

Как получить имяя реворд ?

- $R(dialog, documents, answer)$ - функция качества ответа

Reranker

Value model

- $R(dialog, documents, answer)$ - функция качества ответа
- $V(dialog, documents)$ - функция качества нашего входа
- Не это ли нам надо ?

Reranker

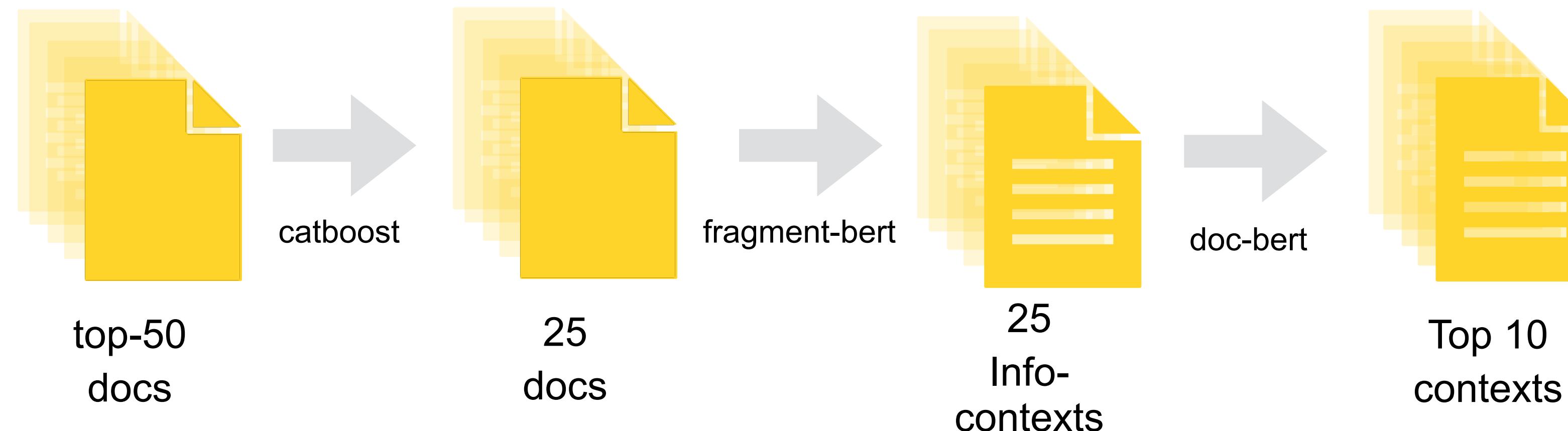
Как получить

- Дистиллируем $E_{answer}R(dialog, documents, answer)$ в $V(dialog, document)$
- Как получить нешумную оценку ?
- Набираем несколько репрезентативных моделей, фиксируем параметры генерации, набираем ответы.
- Также нужно предварительно дообучить модель давать нормальные ответы на одном документе(у нас это был непривычный сценарий)

Reranker

Чуть чуть быстрее

- Дистилируем Bert в catboost на поисковых признаках



ускорение

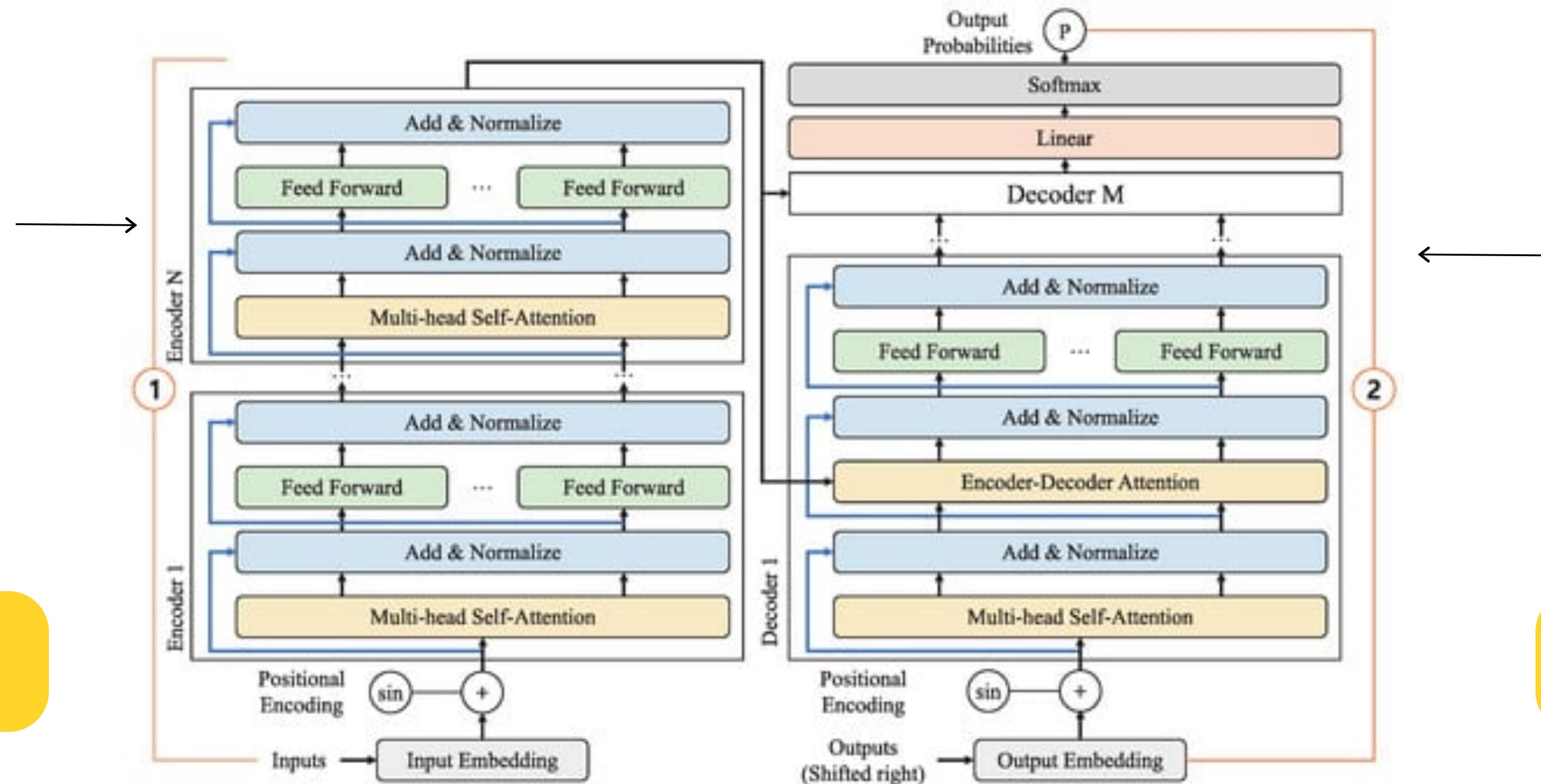
Encoder Decoder

Прогоняется
один раз

2X Слоев

Самая медленная
часть генерации

X Слоев



Generation

Reminder

- Как устроена greedy генерация:
 1. делаем forward pass
 2. добавляем в префикс argmax
 3. повторить пока не сгенерим EOS

Generation

H100 specification

- Перемножения в декодинге имеют вид matrix x vector, поэтому у них сложность $O(n^2 * \text{batch_size})$, а не $O(n^3 * \text{batch_size})$ как в претрейне/контексте
- Из памяти надо грузить слой $O(n^2)$
- Грузить из памяти можем в $\sim 10^3$ раз меньше чем умножать

	H100 SXM	H100 NVL
FP64	34 teraFLOPS	30 teraFLOPs
FP64 Tensor Core	67 teraFLOPS	60 teraFLOPs
FP32	67 teraFLOPS	60 teraFLOPs
TF32 Tensor Core*	989 teraFLOPS	835 teraFLOPs
BFLOAT16 Tensor Core*	1,979 teraFLOPS	1,671 teraFLOPS
FP16 Tensor Core*	1,979 teraFLOPS	1,671 teraFLOPS
FP8 Tensor Core*	3,958 teraFLOPS	3,341 teraFLOPS
INT8 Tensor Core*	3,958 TOPS	3,341 TOPS
GPU Memory	80GB	94GB
GPU Memory Bandwidth	3.35TB/s	3.9TB/s
Decoders	7 NVDEC 7 JPEG	7 NVDEC 7 JPEG

Annotations:

- A green arrow points from the value $4 * 10^{15}$ to the FP8 Tensor Core entry.
- A red arrow points from the value $3 * 10^{12}$ to the GPU Memory Bandwidth entry.

Квантизация

Обзор

Memory bound

- большая модель/небольшой *batch_size*
- ускоряем трансфер VRAM -> shared memory/регистры
- Weight-only методы: GPT-Q, AWQ, AQLM...
- очень дорого, не потянем большой поток

Compute bound

- <=5В модель/большой *batch_size*
- GeMM/V в низкой битности кратно быстрее
- WA методы: SmoothQuant, FP8, QUIK, QuaRot
- можно катить на большой поток

Speculative Decoding

Идея

- Как устроена greedy генерация:
 - делаем forward pass, добавляем в префикс argmax, повторить
- Проблема:
 - меньше batch size => меньше утилизация GPU
- Можно проверить ответ за один forward с лучшей утилизацией
- Идея - пробуем угадать, что сгенерирует модель:
 - повезло => добавили больше одного токена
 - не повезло => лишний compute

<https://arxiv.org/pdf/2302.01318>

Speculative Decoding

Иллюстрация сложных и лёгких токенов

What is the best way to learn a new language?

There is no one-size-fits-all answer to this question, as the best way to learn a new language will depend on your individual learning style, goals, and resources. However, here are some general tips that can help you learn a new language effectively:

1. Set specific goals: Decide what you want to achieve by learning the language, whether it's for travel, work, or personal enrichment. This will help you stay motivated and focused.
2. Immerse yourself: Surround yourself with the language you want to learn, whether it's by listening to music, watching TV shows or movies, or talking to native speakers. This will help you pick up the language more quickly.
3. Practice regularly: Consistency is key when it comes to learning a new language. Set aside dedicated time each day to practice and review what you've learned.
4. Use Learn with AI: AI-powered language learning apps like Duolingo and Rosetta Stone can provide a structured and interactive way to learn new vocabulary, grammar, and sentence structure.
5. Focus on conversation: Practice speaking and listening as much as possible, as this will help

Speculative Decoding

Метод

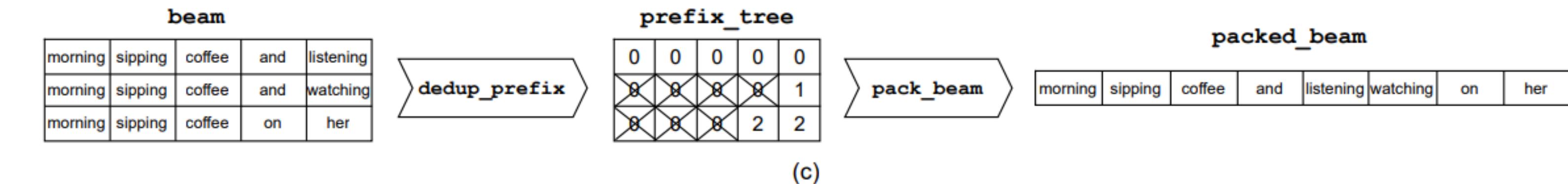
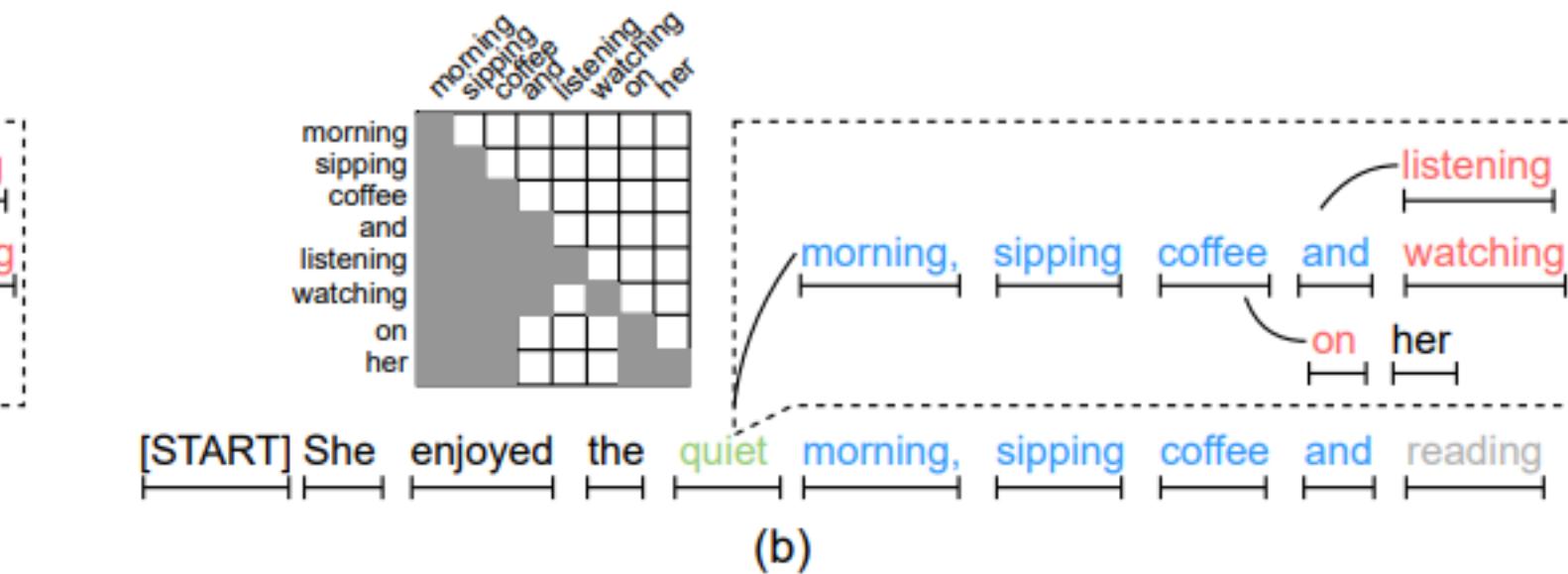
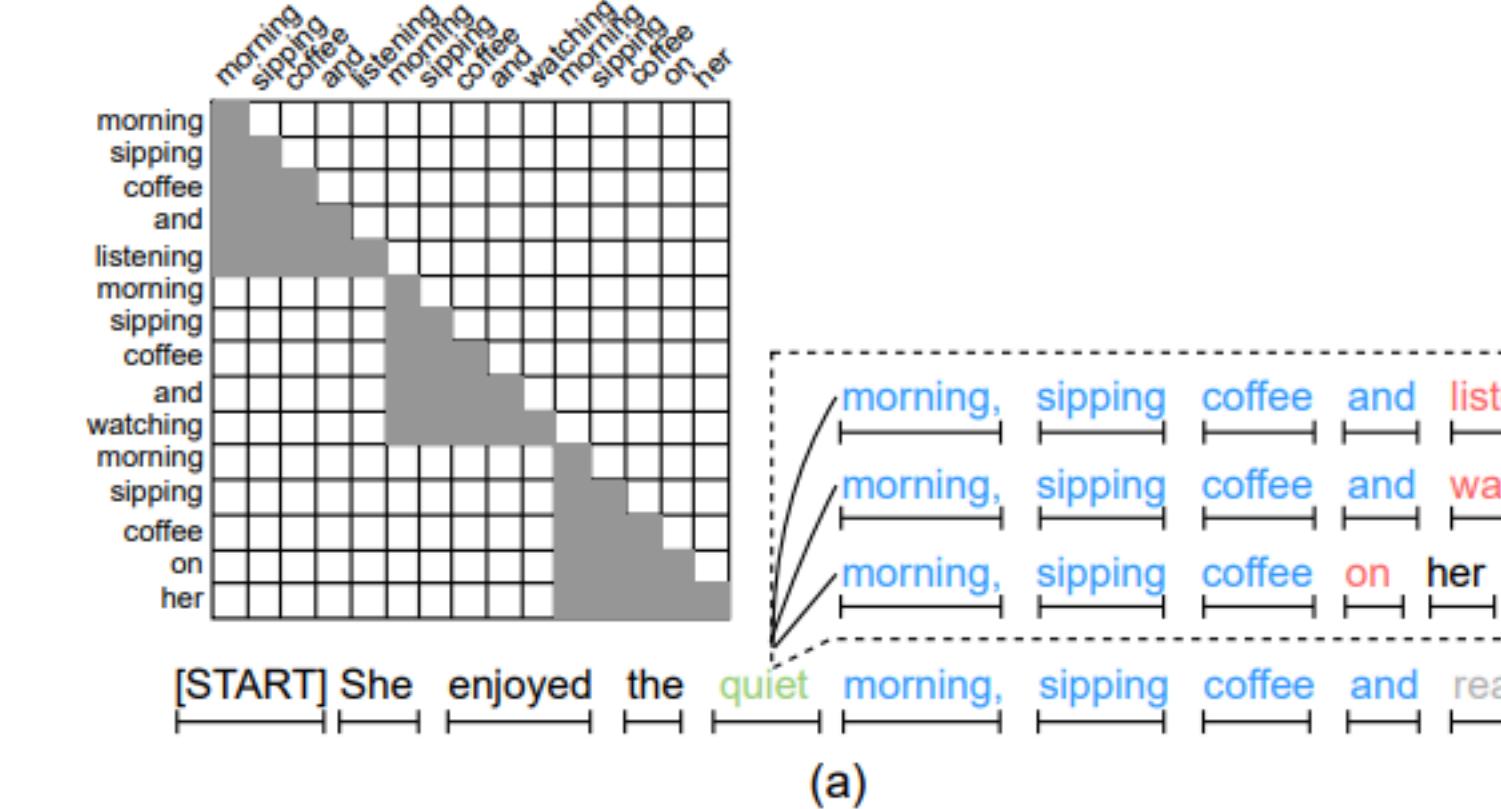
- Есть легковесная процедура drafting-а токенов
- На каждом шаге делаем forward pass нового токена и draft-токенов
- Часть draft-токенов принимается и добавляется в контекст модели
- С точки зрения математики это та же генерация, никаких потерь качества!

Speculative Decoding

ReDrafter

- Метод – однослойная RNN поверх hidden-ов базовой модели

- Тушка заморожена
- Большой AccRate



<https://arxiv.org/pdf/2403.09919>

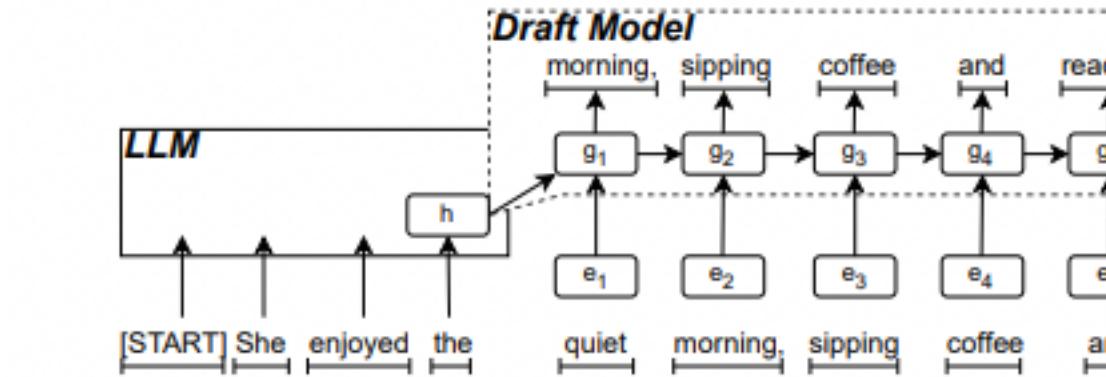


Figure 3: Draft model takes LLM’s last hidden state h as input to forecast next few tokens. For brevity, we omit model parameters and LLM hidden states before h .

Спасибо за внимание и понимание