Начнем с результатов. Они странные. Доля правильных ответов для greedy декодинга получилась самой высокой - 9%. Top-k, top-p, beam search показали себя хуже - 7.2, 7.9, 5.9 процентов соответственно. В принципе, такие результаты соотносятся с пометкой авторов на странице модельки на hf “используйте greedy decoding” для математических задачек, но я думал, что скоры будут все же повыше.
Подход брать самый встречаемый ответ среди ответов, полученных с помощью разных сэмплингов, работает, но результаты различаются незначительно - на 0.4%. Возможно, дело в том, что было перебрано мало комбинаций параметров.

Получилось: наконец-то получилось запустить модель через апишку hf и генерировать ответы на вменяемое время в сравнении с petals.
Не получилось: так как все же апишка имеет ограничения по количеству генераций, то генерация останавливалась и довольно часто приходилось перезапускать, что было не очень приятно. Из-за этого получилось сгенерировать не очень много данных, поэтому я не увидел сильной разницы между CoT и ансамблированным CoT.

Хотелось бы сделать:
1)попробовать разное количество few-shot примеров в промпте(брал столько, сколько брали авторы статьи)
2)перебрать больше комбинаций параметров, сэмплингов
3)можно попробовать исправлять ответ промежуточного действия, потому что я заметил, что бывает действие написано правильно, а ответ посчитан неверно