



Стилизация текста под авторов русской классической литературы

Выбор модели

Попробовали следующие модели:

- ai-forever/rugpt3small_based_on_gpt2
- ai-forever/rugpt3medium_based_on_gpt2
- Gigachat
- Gigachat 2 max
- Yandex GPT-5 lite instruct



Лучший результат показал Gigachat, но из-за нежелания платить за токены выбор пал на Yandex GPT-5, которая тоже демонстрировала хороший результат.

5 000 000 tokens
GigaChat Lite
Сверхскорость для решения задач

1 000 ₽
5,000,000 tokens / year

Buy
У нас проблемы...

336,47 ₽ 10 ₽

Black

Про yandex gpt-5 lite instruct:

YandexGPT 5 Lite на 8B параметров с длиной контекста 32k токенов

Имеет структуру Llama

```
LlamaForCausalLM(  
  (model): LlamaModel(  
    (embed_tokens): Embedding(129024, 4096)  
    (layers): ModuleList(  
      (0-31): 32 x LlamaDecoderLayer(  
        (self_attn): LlamaAttention(  
          (q_proj): Linear8bitLt(in_features=4096, out_features=4096,  
bias=False)  
          (k_proj): Linear8bitLt(in_features=4096, out_features=1024,  
bias=False)  
          (v_proj): Linear8bitLt(in_features=4096, out_features=1024,  
bias=False)  
          (o_proj): Linear8bitLt(in_features=4096, out_features=4096,  
bias=False)  
        )  
        (mlp): LlamaMLP(  
          (gate_proj): Linear8bitLt(in_features=4096,  
out_features=14336, bias=False)  
          (up_proj): Linear8bitLt(in_features=4096, out_features=14336,  
bias=False)  
          (down_proj): Linear8bitLt(in_features=14336,  
out_features=4096, bias=False)  
          (act_fn): SiLU()  
        )  
        (input_layernorm): LlamaRMSNorm((4096,), eps=1e-06)  
        (post_attention_layernorm): LlamaRMSNorm((4096,), eps=1e-06)  
      )  
    )  
    (norm): LlamaRMSNorm((4096,), eps=1e-06)  
    (rotary_emb): LlamaRotaryEmbedding()  
  )  
  (lm_head): Linear(in_features=4096, out_features=129024,  
bias=False)  
)
```

```
LlamaConfig {  
  "architectures": [  
    "LlamaForCausalLM"  
  ],  
  "attention_bias": false,  
  "attention_dropout": 0.0,  
  "bos_token_id": 1,  
  "eos_token_id": 2,  
  "head_dim": 128,  
  "hidden_act": "silu",  
  "hidden_size": 4096,  
  "initializer_range": 0.02,  
  "intermediate_size": 14336,  
  "max_position_embeddings": 32768,  
  "mlp_bias": false,  
  "model_type": "llama",  
  "num_attention_heads": 32,  
  "num_hidden_layers": 32,  
  "num_key_value_heads": 8,  
  "pretraining_tp": 1,  
  "quantization_config": {  
    "_load_in_4bit": false,  
    "_load_in_8bit": true,  
    "bnb_4bit_compute_dtype": "float32",  
    "bnb_4bit_quant_storage": "uint8",  
    "bnb_4bit_quant_type": "fp4",  
    "bnb_4bit_use_double_quant": false,  
    "llm_int8_enable_fp32_cpu_offload": true,  
    "llm_int8_has_fp16_weight": false,  
    "llm_int8_skip_modules": null,  
    "llm_int8_threshold": 6.0,  
    "load_in_4bit": false,  
    "load_in_8bit": true,  
    "quant_method": "bitsandbytes"  
  },  
  "rms_norm_eps": 1e-06,  
  "rope_scaling": null,  
  "rope_theta": 500000.0,  
  "tie_word_embeddings": false,  
  "torch_dtype": "float16",  
  "transformers_version": "4.52.4",  
  "use_cache": true,  
  "vocab_size": 129024  
}
```

Промпты:

Zero-shot

Фамилия автора
Достоевский

Результат
В этот день пробудившись весьма рано от объятий Морфея, воспрянул он духом своим и предался прогулке по окрестностям.

Фамилия автора
Толстой

Результат
В сей день пробудился я до света и отправился на вылазку.

Введите текст
Сегодня я проснулся рано и пошёл на прогулку.

One-shot

Фамилия автора
Достоевский

Результат
"В этот день, ещё до восхода солнца пробудившись от сна своего, вышел я из дому моего с трепетом и волнением великим в сердце моём."

Фамилия автора
Толстой

Результат
"В сей день пробудился я ранее обыкновенного срока и отправился на прогулку пешком."

Промпты:

Few-shot

Результат уже улучшился

Фамилия автора
Гоголь

Результат

"Нет дня, который бы прошёл для меня даром и чтоб я в этот самый день встал поздно! Вот сегодня утром пробудившись раньше всех птиц небесных, немедленно пустился я к прогулке!"

Фамилия автора
Тургенев

Результат

"Сегодня поутру меня разбудила заря. Я встал пораньше и отправился на прогулку по окрестным местам."

Промпты:

One-shot (большой пример текста автора)

Фамилия автора
Достоевский

Результат
Ещё затемно пробудился мой разум от сна и поднял меня, ибо желал я очистить свою головушку утренней прогулкой. И вот вышел я робко, словно тень бесплотнейшая, дабы никто и ничто не потревожили мои думы глубокие да невесёлые. Шел я неспешно, озираясь кругом и примечая всё вокруг себя, словно боялся пропустить нечто важное, некую истину сокрытую.

Фамилия автора
Пушкин

Результат
Утро раннее разбудило меня. С сердцем веселым встал я спозаранку, пошел гулять один, под небо чистое гляючи да воздух свежий вдыхаючи.

Few-shot (большой пример текста автора + примеры)

В комбинации примеры + большой текст автора получается добиться лучших результатов

Фамилия автора
Достоевский

Результат
Едва открыв глаза поутру и чувствуя необъяснимую лёгкость во всём теле, я встал с намерением совершить небольшую прогулку по улицам сего города, столь мрачного и унылого с первого взгляда, однако полного тайн и загадок, коими он непрестанно меня манит. Улицы пробудили во мне неясные мысли, полные тревоги и волнения, сердце моё билось чаще, словно предчувствуя нечто великое и ужасное одновременно. Я шёл быстрым шагом, погружённый в свои размышления, и чувствовал, как воздух наполняется новыми ароматами жизни и неведомыми возможностями.

Фамилия автора
Пушкин

Результат
Заря уж давно осветила наши кровли, а я меж тем пробудился от сна моего и встал поутру прогуляться. Сердце моё наполнилось утренней свежестью, и душа возжелала уединения среди природы. И вот выхожу я – гляжу вокруг себя да вдыхаю воздух свежий полной грудью...

Метрика:



Построена собственная нейросетевая метрика

- На основе DeepPavlov
- Обучена на 9 авторах, для выявления их стилей
- Результат оставляет желать лучшего

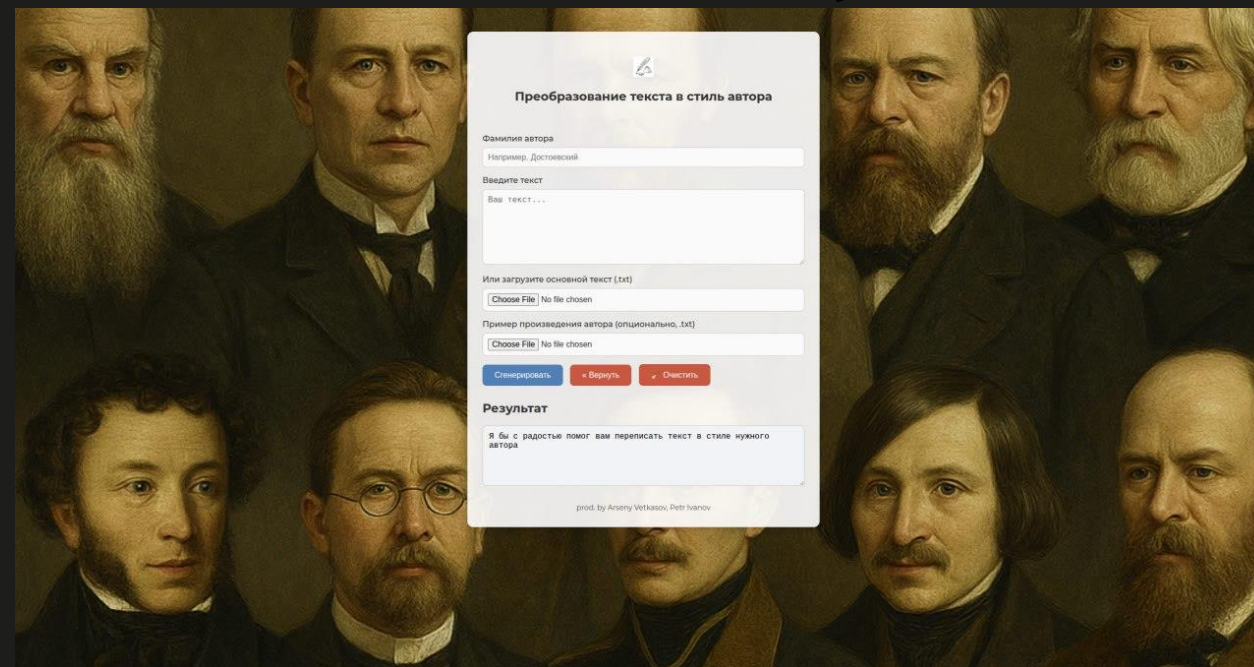
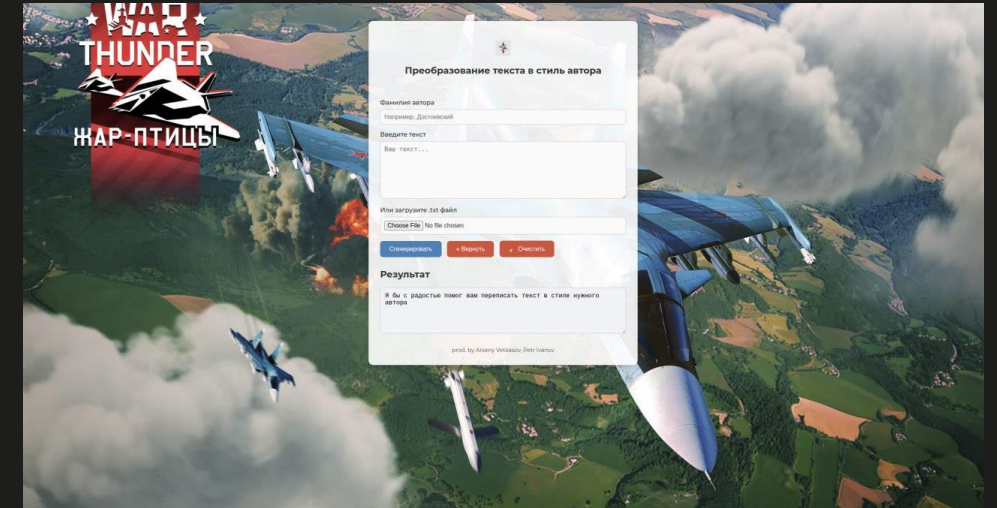
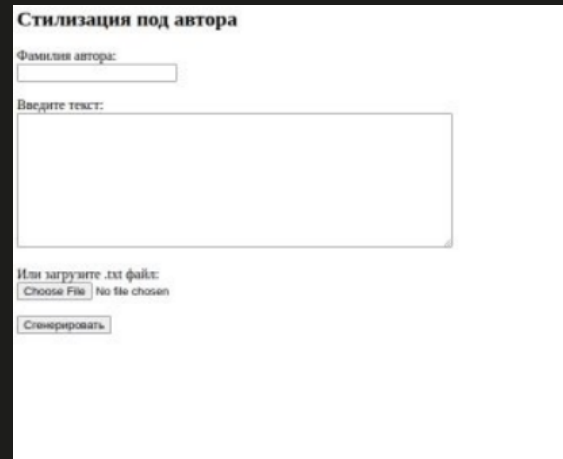
Причины:

- 1) Малый объём данных
- 2) Специфика обучения
- 3) Недостаточная мощность модели

Построение интерфейса:

Интерфейс на основе flask

Был создан скрипт, который запускает приложение (сайт на html) и обращается непосредственно к модели для получения результатов генерации.

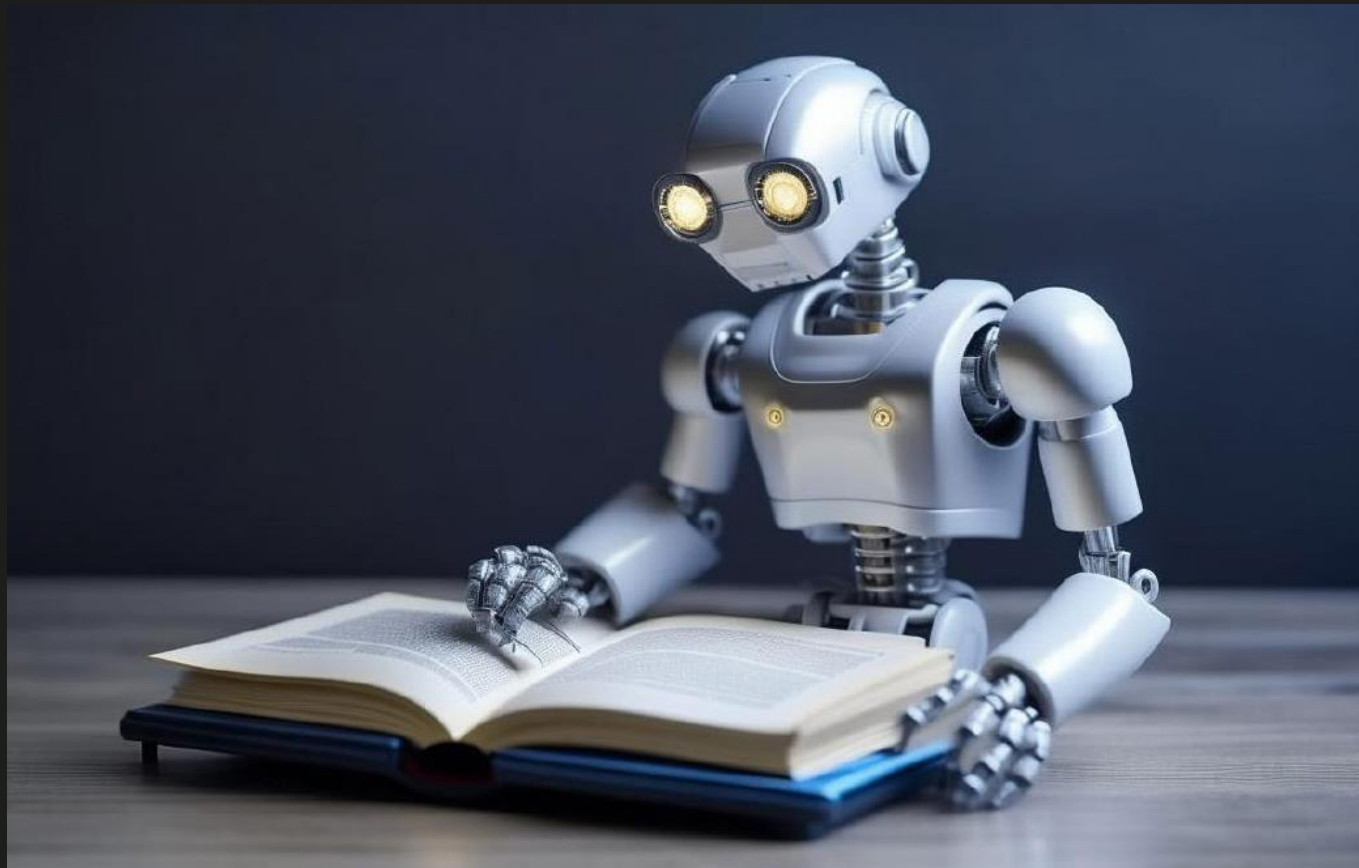


Эволюция:

Дообучение:

Fine-tuning LoRA:

- С Использованием мощностей Kaggle
- Запуск зафайнтюненной модели локально
- Небольшое улучшение генерации



Проблемы:

- 1) Ограничения портала, сложность выгрузки весов
- 2) Специфика обучения
- 3) Малый объём данных

Итоговая версия проекта:

Основа в виде модели

Yandex GPT-5 lite instruct

Подобранный промпт

Используем few-shot с примером текста автора

Визуальный интерфейс

Flask + html

Альтернативный вариант

Fine-tuning методом LoRA

