API Current Specs

Key value pair definitions:

API accepts json. Transmit key descriptions below. Transmit all keys on each instance. Desired value is not know then set equal to default value.

- "api_key" = "dW8tB$j3yx&KvEvsP8QSt24&M2%QwYXD" Key to access the api. If not set the api will return string "Invalid api_key."
- "pred_name" = "model-size-genre-steps", e.g. "onnx-quant-774M-Pulitzer-30ksteps". This parameter is so that you can access multiple models at a single endpoint. default="onnx-quant-774M-Pulitzer-30ksteps"
- "mode" = "s-completion", "s-completion+", or "p-completion". It is the mode to be used in. "s-completion", completes the sentence. "s-completion+", completes the sentence and writes another. "p-completion" completes the paragraph.
- "text" = "Put text here". Don't put more than 3500 characters. This is the text accepted to generate the completion
- "temperature" = decimal between 0 and 2, default = .7
- "repetition_penalty" = decimal, penalty from Salesfor Control paper, default = 1.2
- "top_k" = integer, 0(off) to 1000, default = 0
- "top_p" = decimal, 0 to 1, default = 1
- "batch_size" = integer, default = 3

API returns json. The return key values are below.

- "text" = {0: "text batch zero", 1: "text batch one", 2: "text batch two"}. This holds a dictionary with the generated text.
- "truncated" = true or false . Tells when the input text was too long. The api automatically truncates.

Sample request

CURL request:

```
curl yourapiendpointgoeshere -X POST -H "Content-Type: application/json" -d @sample.json
```

Sample.json file contents:

{"api_key": "dW8tB$j3yx&KvEvsP8QSt24&M2%QwYXD", "pred_name": "model1", "mode": "s-completion", "text": "As I walk through the house, i feel scared. There is something not right, something not right here. I heard", "temperature": 0.8, "repetition_penalty": 1.0, "top_k": 0, "top_p": 0.9, "batch_size": 2}

Notes for future development

Thumb rules for calculations

- 1.5 tokens per word.
- Average word is 5 characters

gen_len = integer, max number how many words to generate in addition to output, tokens for time being, Sentence completion(30 words, 45 tokens), sentence completion + another(60 words, 90 tokens), paragraph completion(300 words, 450 tokens), included in the mode