



Technique2: Faster Constrained Decoding

Liangsheng Yin

Shanghai Jiao Tong University & UC Berkeley



Constrained Decoding Matters



OpenAI JSON Mode



Copilot Code Generation



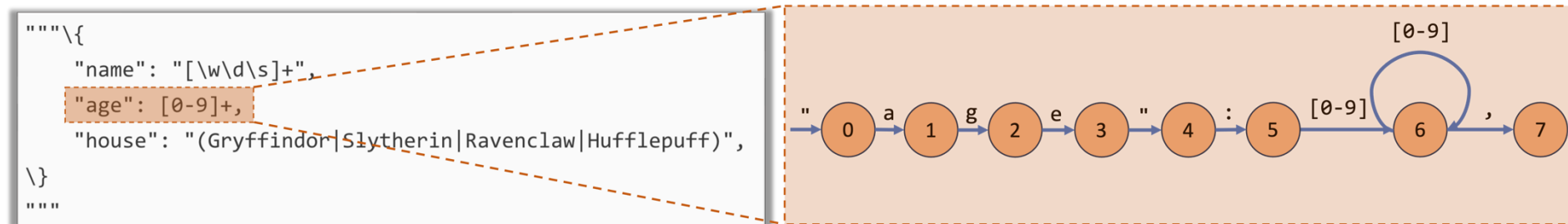
LangChain

LangChain JSON parser

- output will always adhere to the specified syntactic constraints
- reduces the need for ad-hoc parsing, retrying, and prompting
- without the need for fine-tuning or additional post-processing

Constrained decoding works by masking the invalid tokens

Constraint decoding: JSON schema -> regular expression -> finite state machine -> logit mask



Regular Expression

Finite State Machine

Please fill in the following information about Harry Potter.
{
 "name": "Harry",
 "

Decoding Status

Decode + FSM



- age ✓
- Age ✗
- hou ✗

Please fill in the following information about Harry Potter.
{
 "name": "Harry",
 "age":

Decoding Status

Decode + FSM



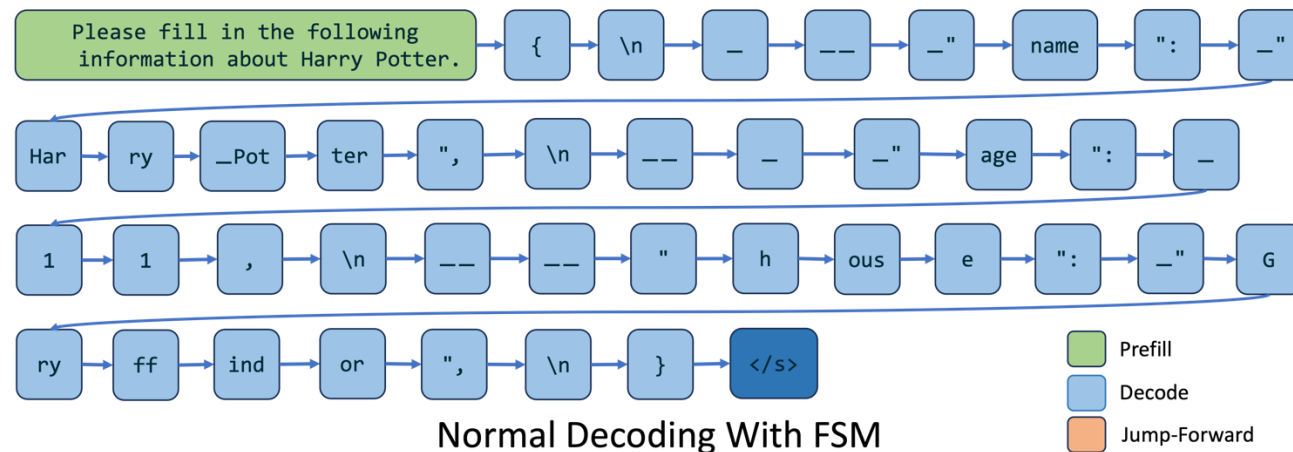
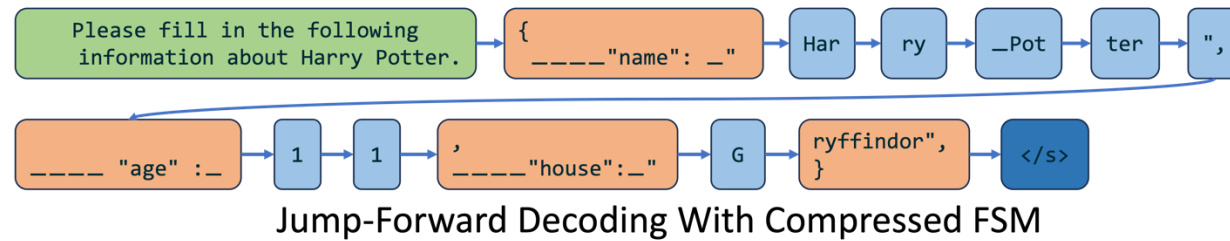
- 0 ✓
- 1 ✓
- fif ✗

✓ allowed next token
✗ not allowed next token

Constrained Decoding With Logits Mask

Compressing the finite state machine allows decoding multiple tokens

We can compress many deterministic paths in the state machine



```
Please fill in the following
information about Harry Potter.
{
  "name": "Harry",
  "age": 15,
  "house": "Gryffindor"
}
```

```
Please fill in the following
information about Harry Potter.
{
  "name": "Harry",
  "age": 15,
  "house": "Gryffindor"
}
```

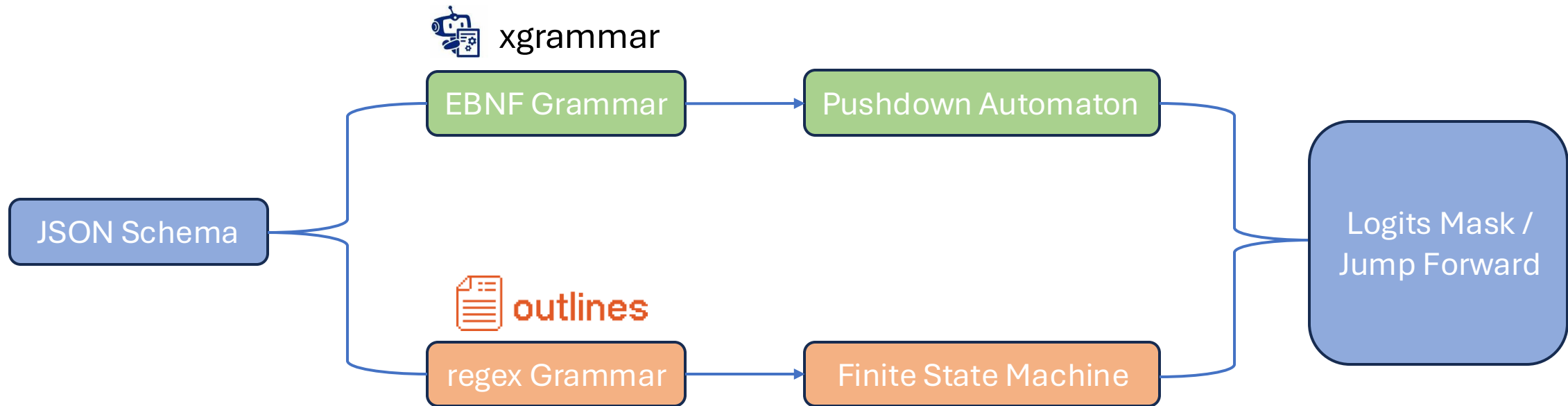
Generated JSONs

Efficient constrained decoding



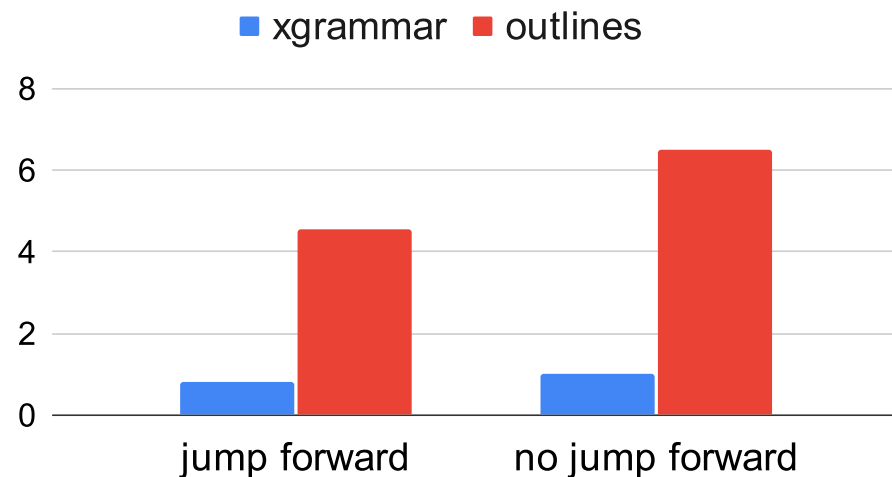
Constrained Decoding Backend: xgrammar and Outlines

We benchmark with 400 JSON schemas from [\(BFCL\) Berkeley Function-Calling Leaderboard](#)

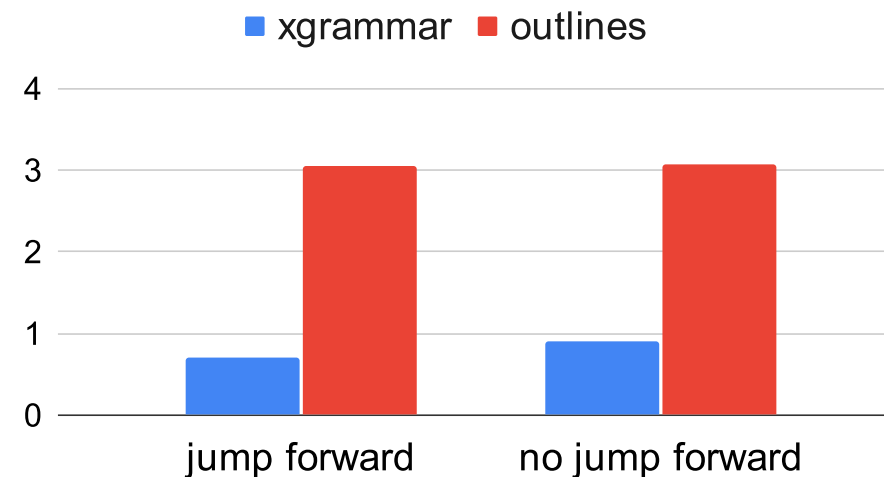


Benchmark Result (Lower is Better)

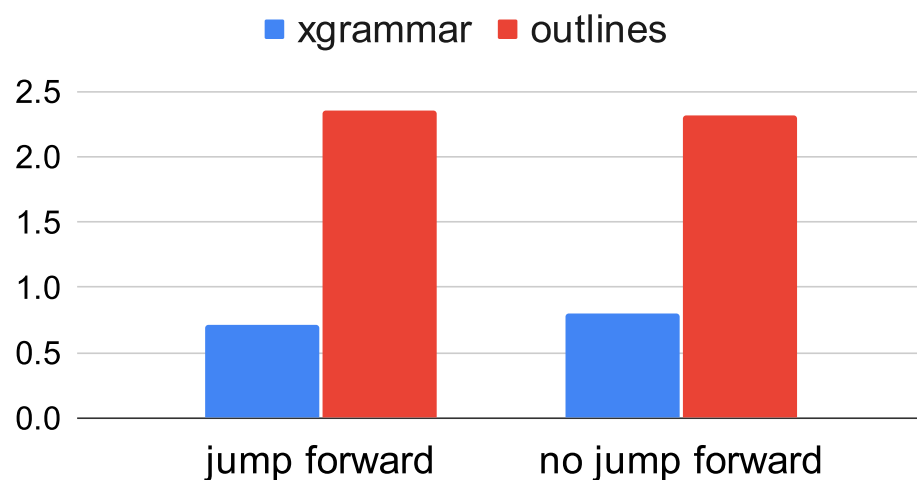
Llama3.1-8B: Single Request Latency (s)



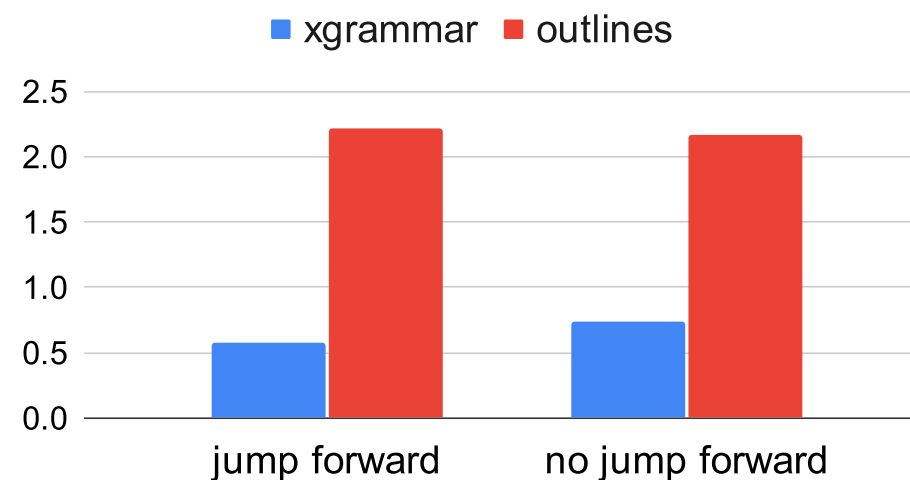
Llama3.1-8B: Batch Request Latency (s)



DeepSeek-V2-Lite: Single Request Latency (s)



DeepSeek-V2-Lite: Batch Request Latency (s)



Question & Answer

Thanks to Ziyi, the xgrammar integration with SGLang is on the way:

<https://github.com/sgl-project/sglang/pull/1680>

Yixin will deliver more details about xgrammar soon!