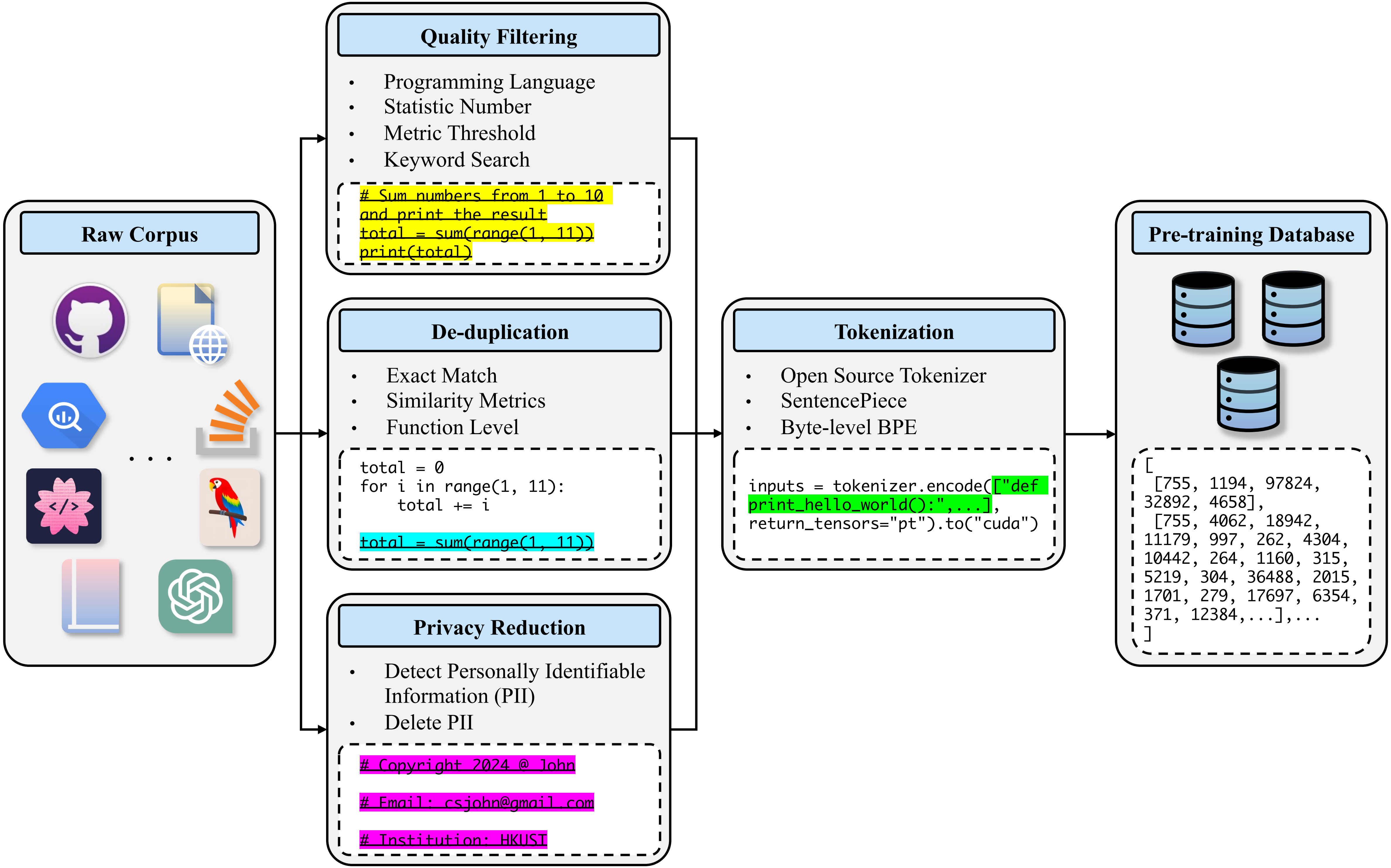

Raw Corpus

Quality Filtering

- Programming Language
- Statistic Number
- Metric Threshold
- Keyword Search

```
# Sum numbers from 1 to 10
and print the result
total = sum(range(1, 11))
print(total)
```

De-duplication

- Exact Match
- Similarity Metrics
- Function Level

```
total = 0
for i in range(1, 11):
    total += i

total = sum(range(1, 11))
```

Privacy Reduction

- Detect Personally Identifiable Information (PII)
- Delete PII

```
# Copyright 2024 @ John

# Email: csjohn@gmail.com

# Institution: HKUST
```

Tokenization

- Open Source Tokenizer
- SentencePiece
- Byte-level BPE

```
inputs = tokenizer.encode(["def
print_hello_world():",...],
return_tensors="pt").to("cuda")
```

Pre-training Database

```
[
 [755, 1194, 97824,
32892, 4658],
 [755, 4062, 18942,
11179, 997, 262, 4304,
10442, 264, 1160, 315,
5219, 304, 36488, 2015,
1701, 279, 17697, 6354,
371, 12384,...],...
]
```