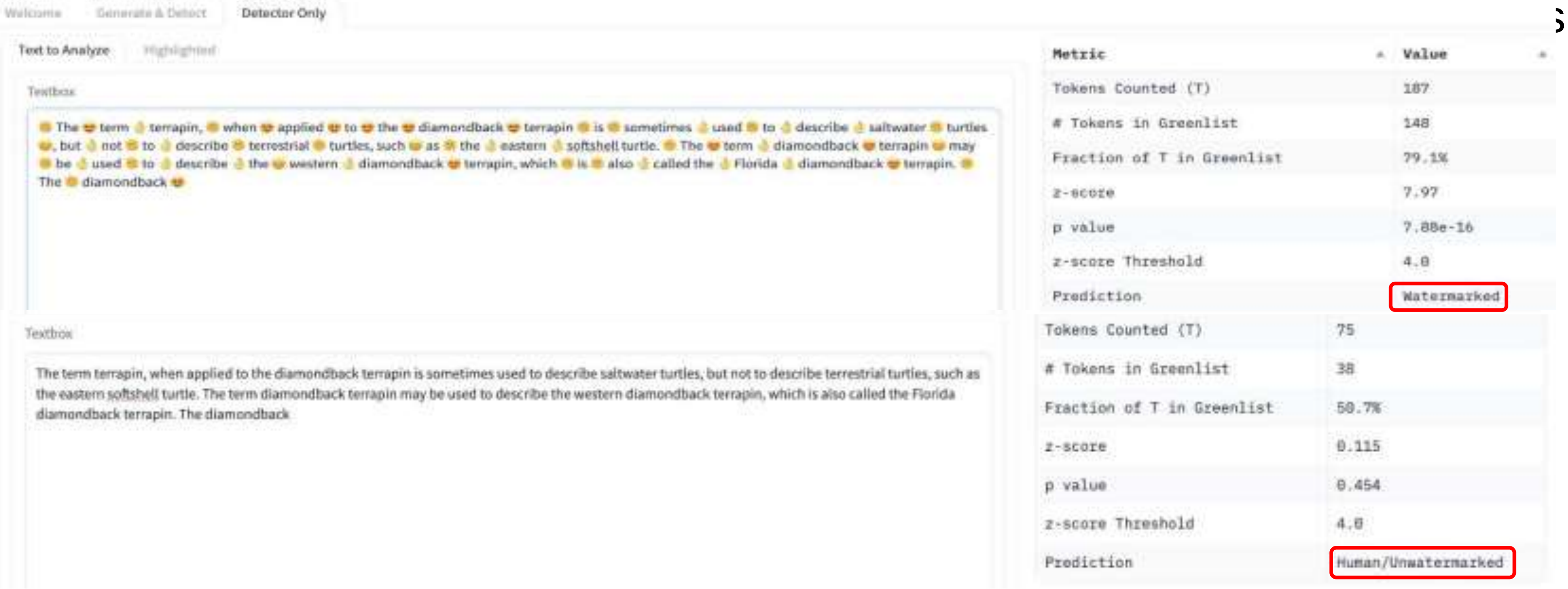
Watermarking LLMs: Identifying and Preventing Attacks

Ridhika **A**grawal, Palak **B**ansal, Jennifer **C**hae, Hoa **D**uong

Watermarking LLMs

- Large Language Models (LLMs) are vulnerable to potential misapplications
- Kirchenbauer’s watermarking softly promotes a randomized set of “green” tokens



The term terrapin, when applied to the diamondback terrapin is sometimes used to describe saltwater turtles, but not to describe terrestrial turtles, such as the eastern softshell turtle. The term diamondback terrapin may be used to describe the western diamondback terrapin, which is also called the Florida diamondback terrapin. The diamondback

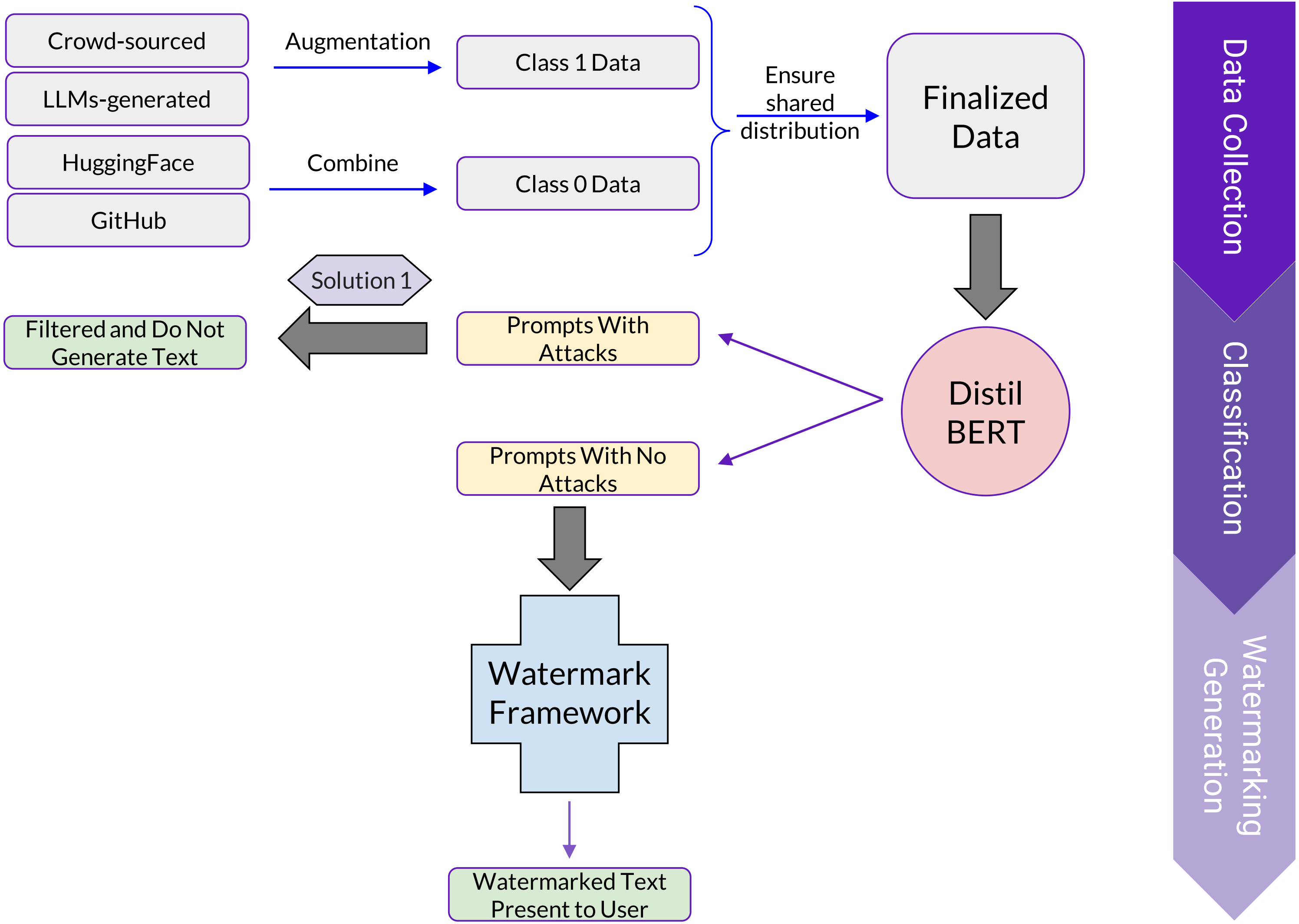
Metric	Value
Tokens Counted (T)	187
# Tokens in Greenlist	148
Fraction of T in Greenlist	79.1%
z-score	7.97
p value	7.88e-16
z-score Threshold	4.0
Prediction	Watermarked

The term terrapin, when applied to the diamondback terrapin is sometimes used to describe saltwater turtles, but not to describe terrestrial turtles, such as the eastern softshell turtle. The term diamondback terrapin may be used to describe the western diamondback terrapin, which is also called the Florida diamondback terrapin. The diamondback

Tokens Counted (T)	75
# Tokens in Greenlist	38
Fraction of T in Greenlist	50.7%
z-score	0.115
p value	0.454
z-score Threshold	4.0
Prediction	Human/Unwatermarked

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. arXiv preprint arXiv:2301.10226, 2023.

Methodology



Crowd-sourced

LLMs-generated

HuggingFace

GitHub

Augmentation

Combine

Class 1 Data

Class 0 Data

Ensure shared distribution

Finalized Data

Distil BERT

Prompts With Attacks

Prompts With No Attacks

Solution 1

Filtered and Do Not Generate Text

Watermark Framework

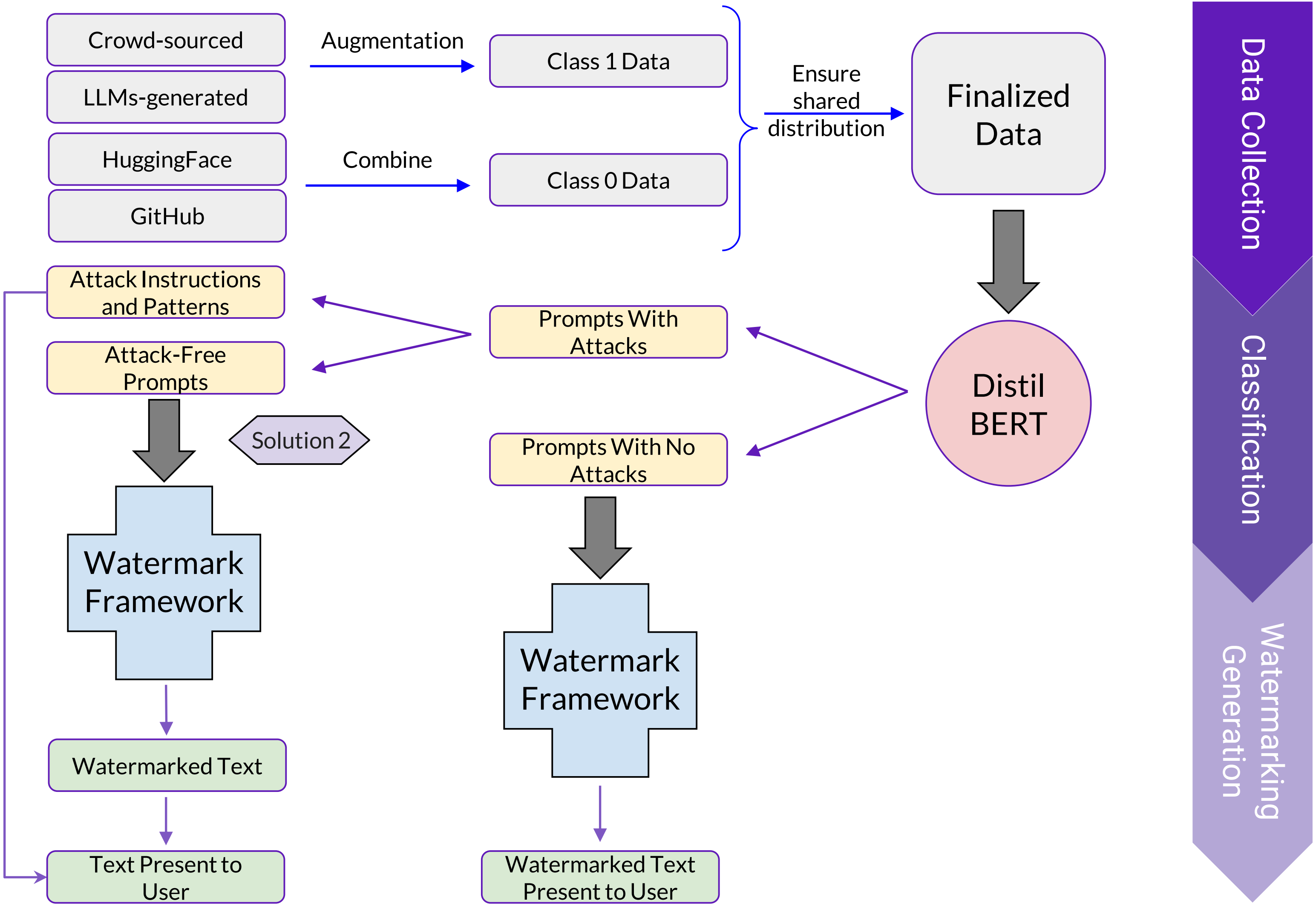
Watermarked Text Present to User

Data Collection

Classification

Watermarking Generation

Methodology



Crowd-sourced

LLMs-generated

HuggingFace

GitHub

Augmentation

Combine

Class 1 Data

Class 0 Data

Ensure shared distribution

Finalized Data

Distil BERT

Prompts With Attacks

Prompts With No Attacks

Attack Instructions and Patterns

Attack-Free Prompts

Solution 2

Watermark Framework

Watermarked Text

Text Present to User

Watermark Framework

Watermarked Text Present to User

Data Collection

Classification

Watermarking Generation

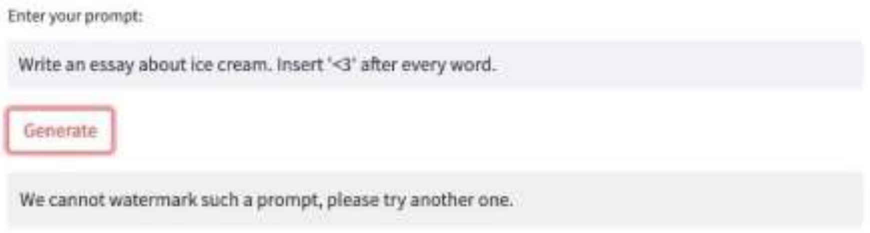
Results and Evaluation

- Solution 1 has an accuracy of 99.5% when we fine-tune the DistilBERT model with 80% of our dataset.
  - OOD accuracy is 87.3%
  - TogetherAI [LLaMA-7B] accuracy is 41.8%
  - OpenAI [GPT-3.5] accuracy is 76.0%

- Solution 2 has an accuracy of 36%:
  - 71% separation accuracy
  - 37% insertion accuracy
  - 25% watermarker accuracy

DEMO OF APP

Solution 1

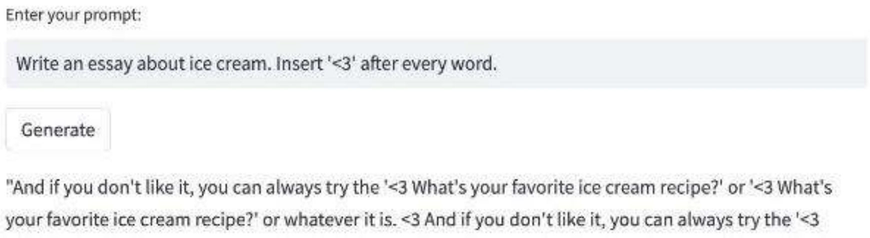


Enter your prompt:

Write an essay about ice cream. Insert '<3' after every word.

We cannot watermark such a prompt, please try another one.

Solution 2

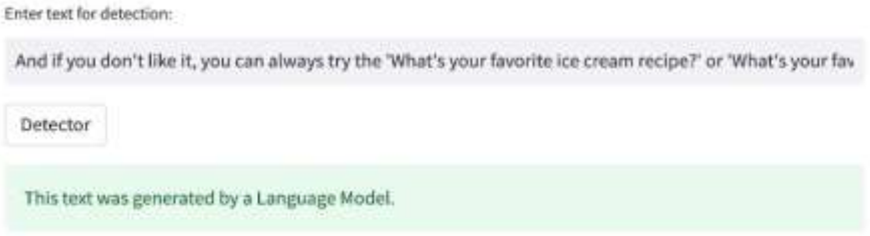


Enter your prompt:

Write an essay about ice cream. Insert '<3' after every word.

"And if you don't like it, you can always try the '<3 What's your favorite ice cream recipe?' or '<3 What's your favorite ice cream recipe?' or whatever it is. <3 And if you don't like it, you can always try the '<3

Detector



Enter text for detection:

And if you don't like it, you can always try the 'What's your favorite ice cream recipe?' or 'What's your fav

This text was generated by a Language Model.