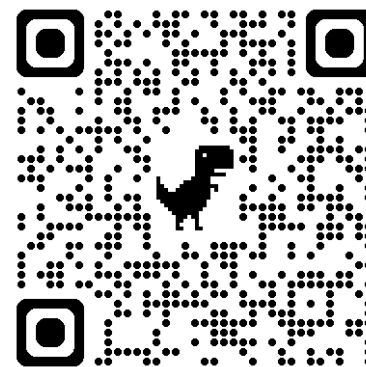


ACL 2024

Bangkok, Thailand



Website; Q&A

Watermarking for Large Language Models

Part V: Conclusion



Xuandong Zhao

UC Berkeley



Yu-Xiang Wang

UC San Diego



Lei Li

CMU

Outline

- Part I: Introduction
 - Part II: Text Watermark
 - (a) Green-Red Watermark
 - (b) Cryptographic Watermark
 - (c) Theoretical results
-
- Part III: Model Watermark
 - Part IV: Post-Hoc Text Detection
 - Part V: Conclusion and Future Directions

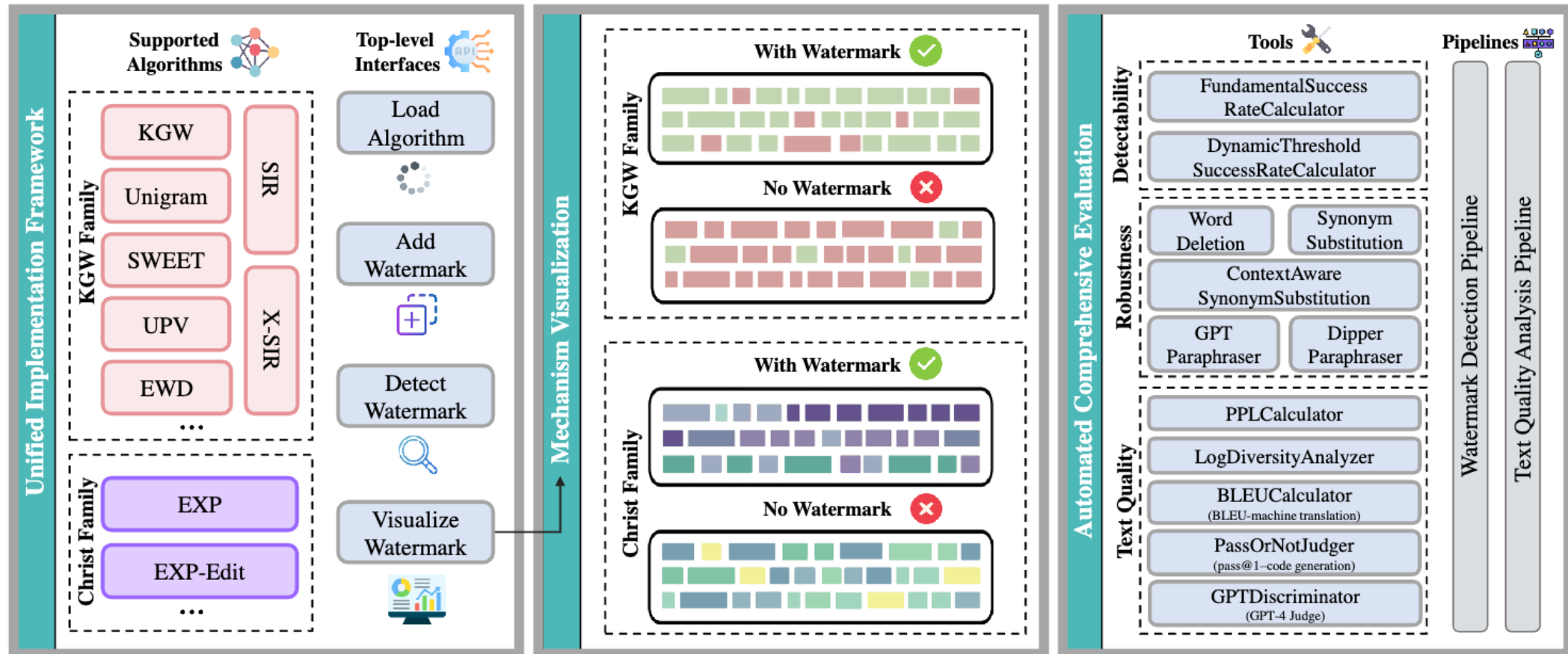
Break

Benchmarks

- Mark My Words: Analyzing and Evaluating Language Model Watermarks
- WaterBench: Towards Holistic Evaluation of Watermarks for Large Language Models (ACL 2024)
- New Evaluation Metrics Capture Quality Degradation due to LLM Watermarking (TMLR 2024)
- WaterJudge: Quality-Detection Trade-off when Watermarking Large Language Models (NAACL 2024)
- ...

Toolkit

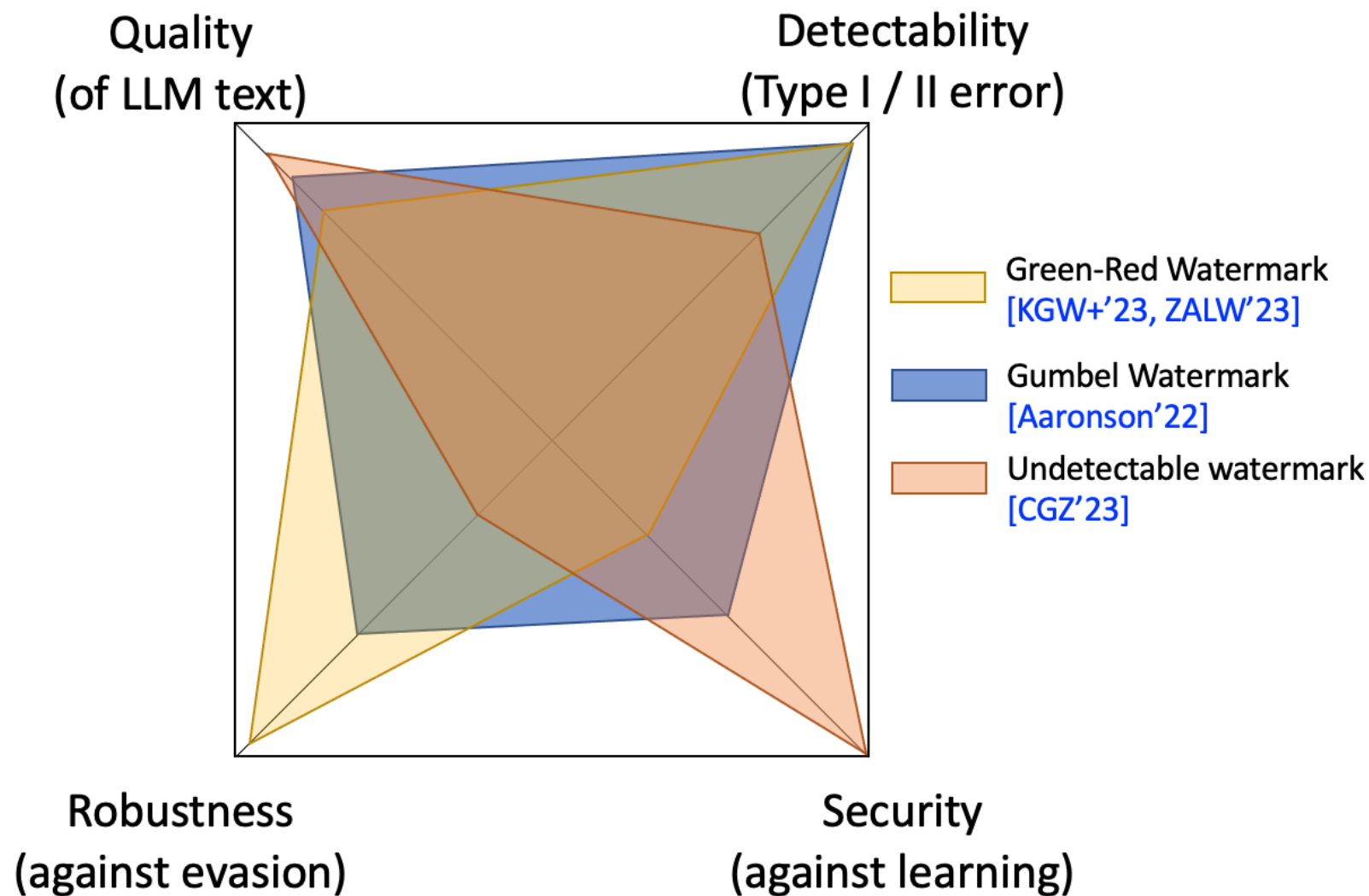
- MarkLLM: An Open-Source Toolkit for LLM Watermarking



Open Problems for Model Watermark

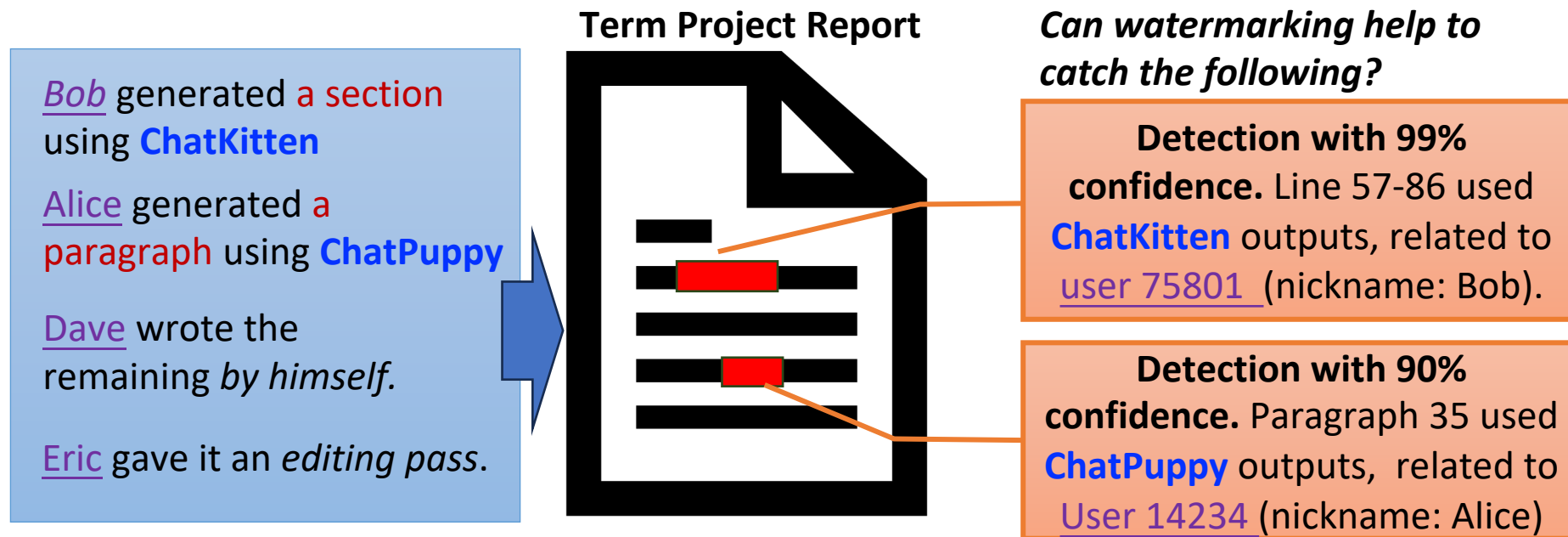
- Defend against multiple/all attacks
 - Distillation
 - Finetuning
 - Pruning
- Theoretical guarantee
 - Quality
 - Detection accuracy
 - Robustness
 - Security

Open Problems: Optimal Tradeoffs



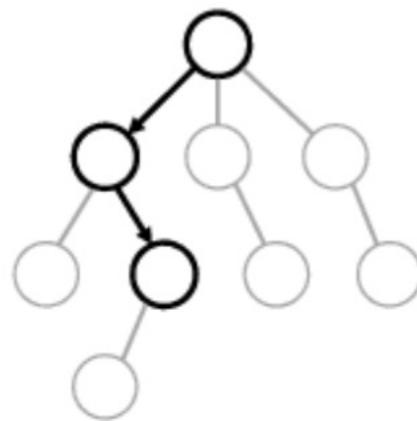
Open Problems: Enhancing Robustness

- Optimality in the edit model. Is Unigram WM the optimal?
- More realistic threat models



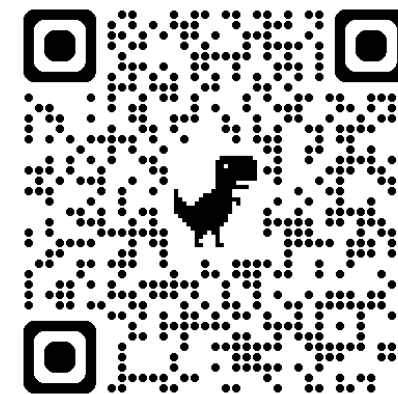
Open Problems: More co-design of decoder and watermarks?

- Provable Watermarking for Beam search?
Or other methods that aim at solving the sequence level MLE decoding.
- When can we still watermark without entropy?



Open Problem: Efficient Evaluation of Model Watermark

- Current setup is inefficient
 - 100 candidate models (50 positive, 50 negative). need to conduct inference for each.
- How to properly conduct comprehensive evaluation of model watermark quality/accuracy/robustness



Thanks for listening!

Questions?

https://leililab.github.io/llm_watermark_tutorial/



Xuandong Zhao
UC Berkeley



Yu-Xiang Wang
UC San Diego



Lei Li
CMU