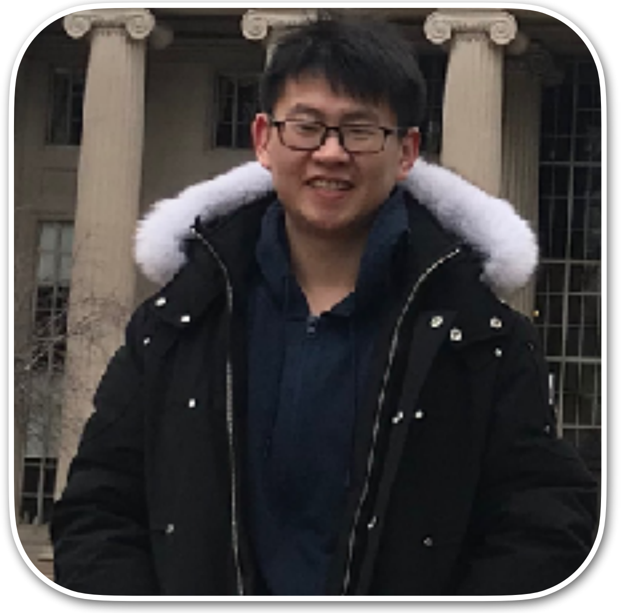
Distributional Information Embedding: A Framework for LLM Watermarking

Haiyun He
Postdoc @ Center for Applied Math, Cornell University

Yepeng Liu
Univ. of Florida



Prof. Ziqiao Wang
Tongji Univ.

Prof. Yongyi Mao
Univ. of Ottawa

Prof. Yuheng Bu
Univ. of Florida

8th C^3 Workshop on Cognition & Control @ University of Florida

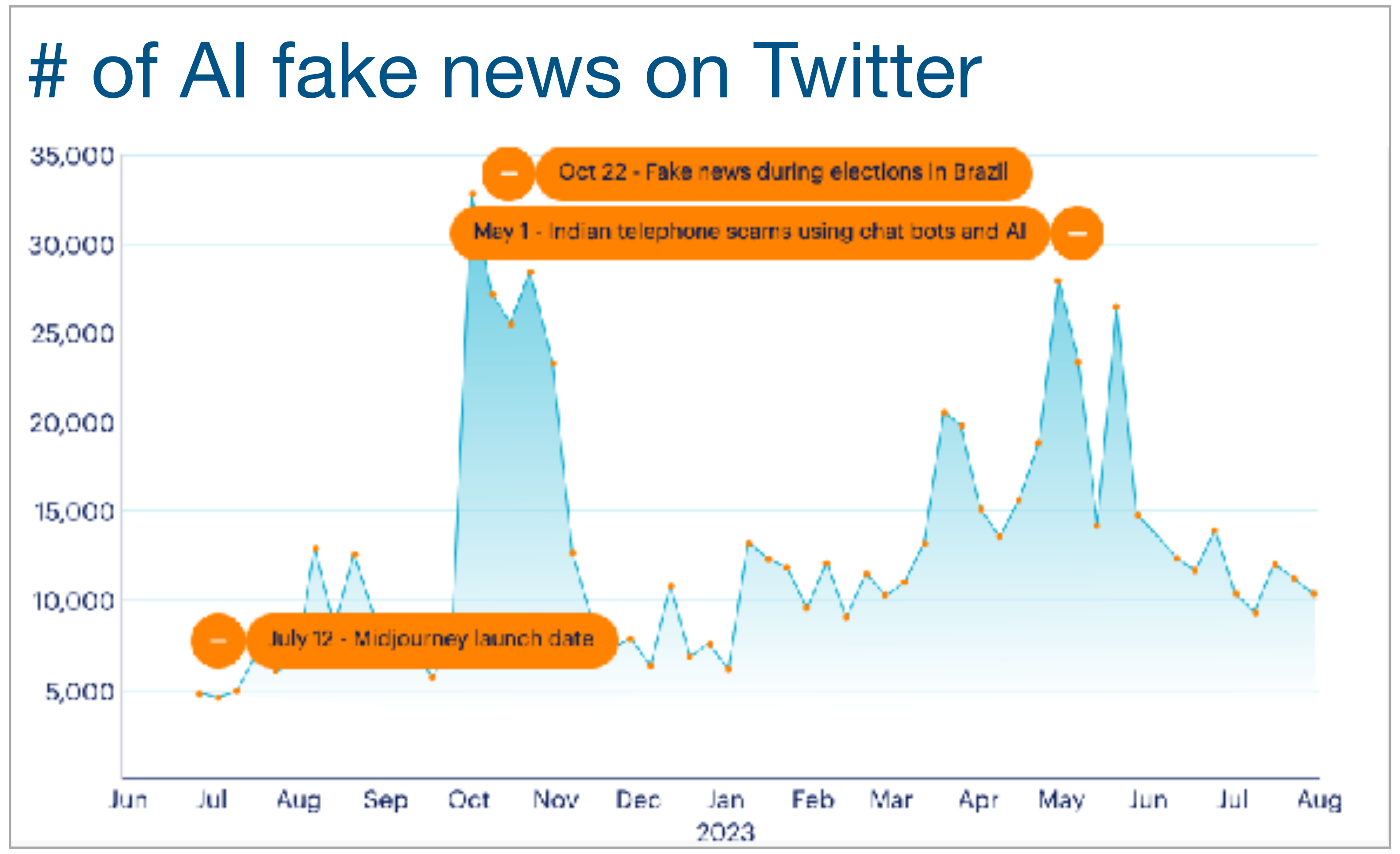
Challenges in AI Safety

Misuse of AI-generated content

[Source] Netskope.com and public sources.

Challenges in AI Safety

Misuse of AI-generated content



# of AI fake news on Twitter

Oct 22 - Fake news during elections in Brazil

May 1 - Indian telephone scams using chat bots and AI

July 12 - Midjourney launch date

Fake news

[Source] Netskope.com and public sources.

Challenges in AI Safety

Misuse of AI-generated content

# of AI fake news on Twitter

AI scams

[Source] Netskope.com and public sources.

Challenges in AI Safety

Misuse of AI-generated content

# of AI fake news on Twitter

Plagiarism

[Source] Netskope.com and public sources.

Challenges in AI Safety

Misuse of AI-generated content

Data Pollution

# of AI fake news on Twitter

Plagiarism

[Source] Netskope.com and public sources.

Challenges in AI Safety

Misuse of AI-generated content

Plagiarism

Data Pollution

Tons of AI-generate data over the internet

[Source] Netskope.com and public sources.

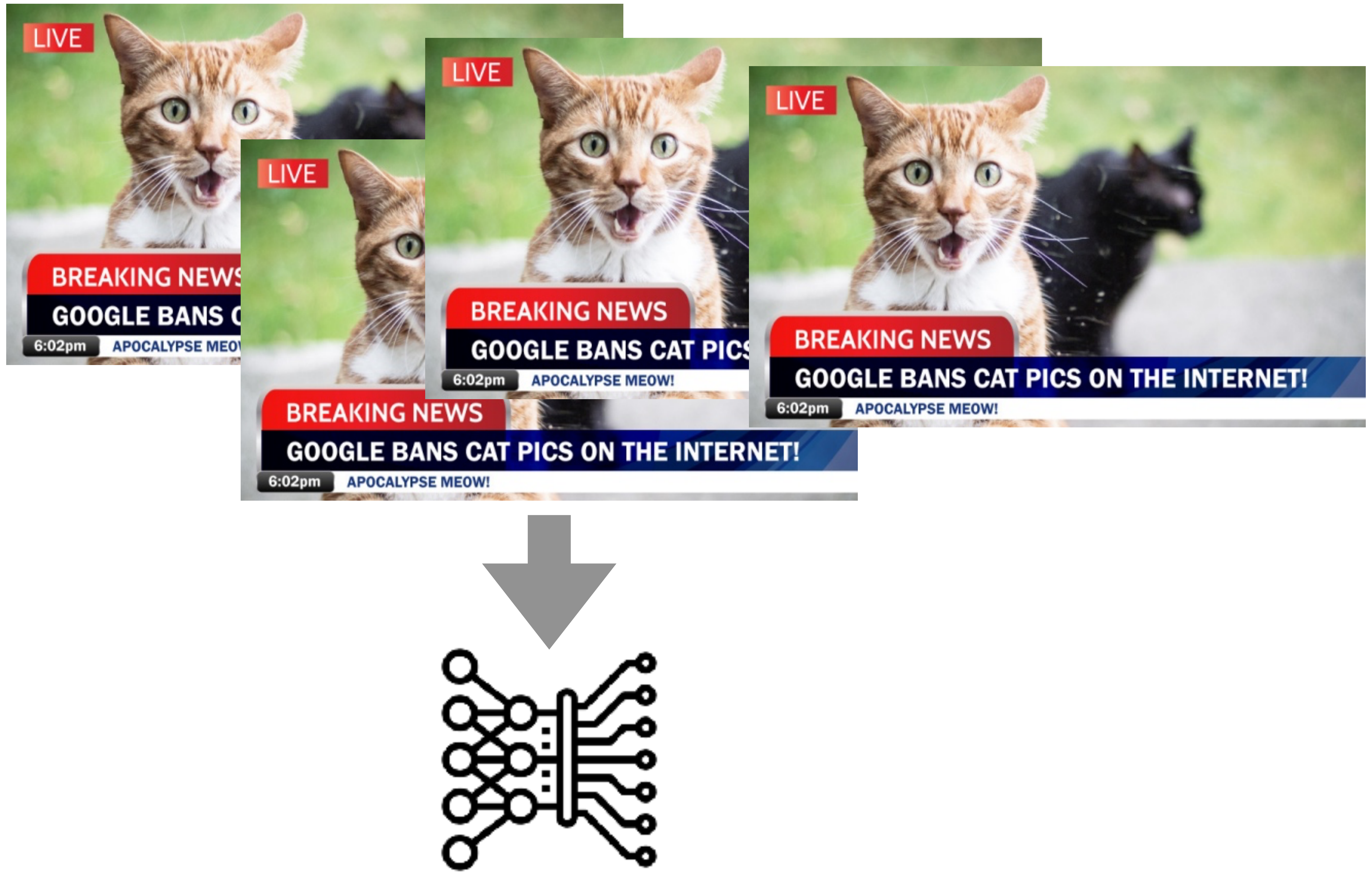
Challenges in AI Safety

Misuse of AI-generated content

Plagiarism

Data Pollution

Tons of AI-generate data over the internet



[Source] Netskope.com and public sources.

Challenges in AI Safety

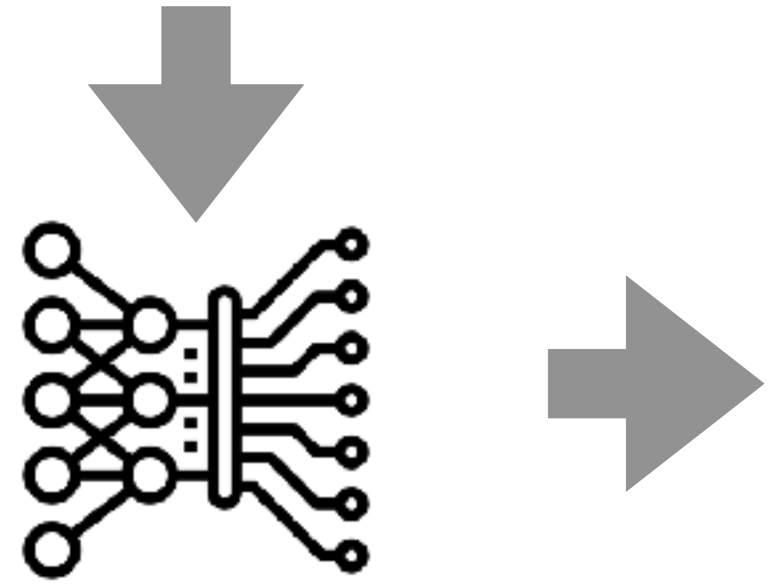
Misuse of AI-generated content

# of AI fake news on Twitter

Plagiarism

Data Pollution

Tons of AI-generate data over the internet



???
No cats anymore?

[Source] Netskope.com and public sources.

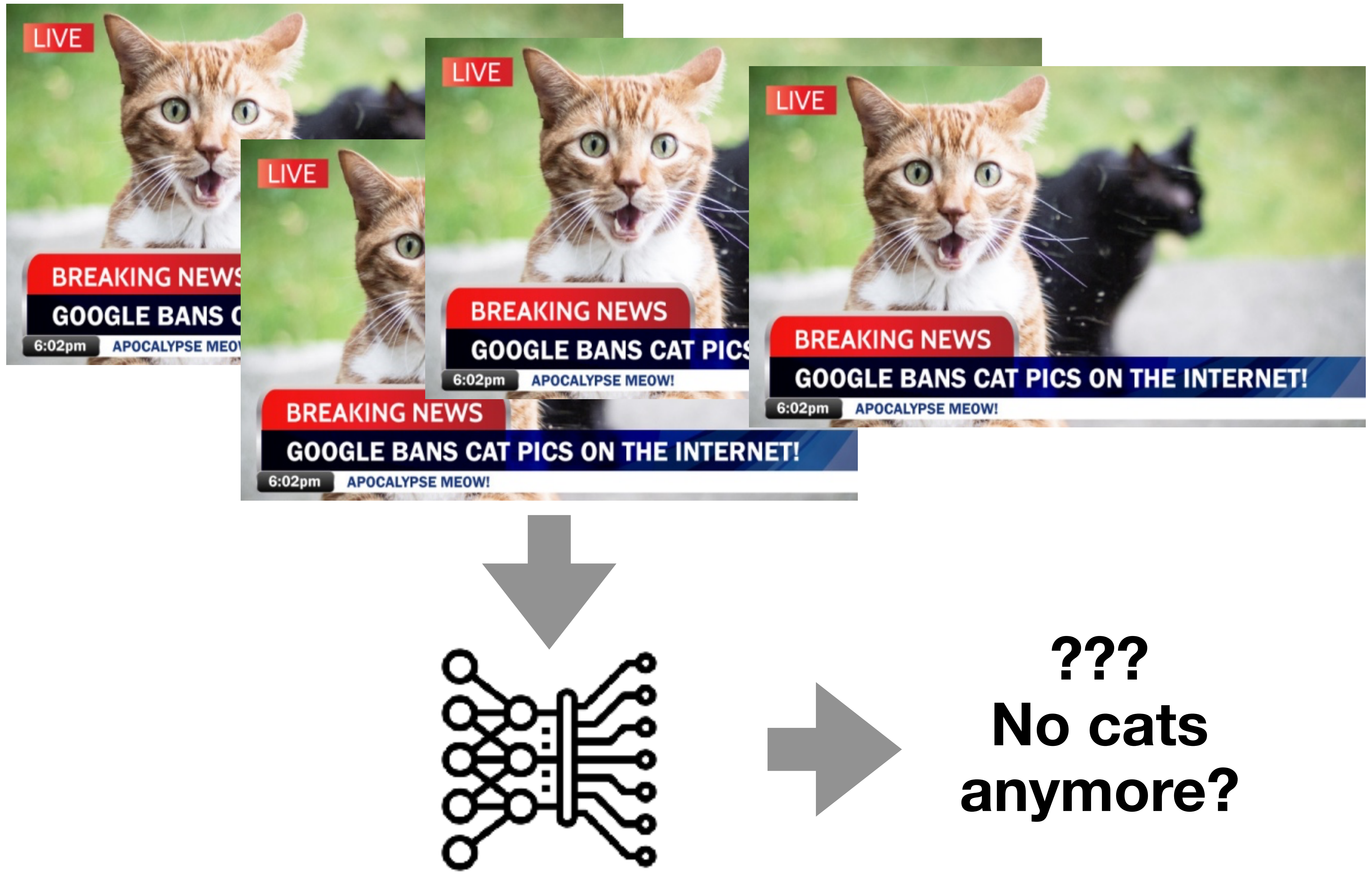
Challenges in AI Safety

Misuse of AI-generated content

Plagiarism

Data Pollution

Tons of AI-generate data over the internet



???
No cats anymore?

collapse…

[Source] Netskope.com and public sources.

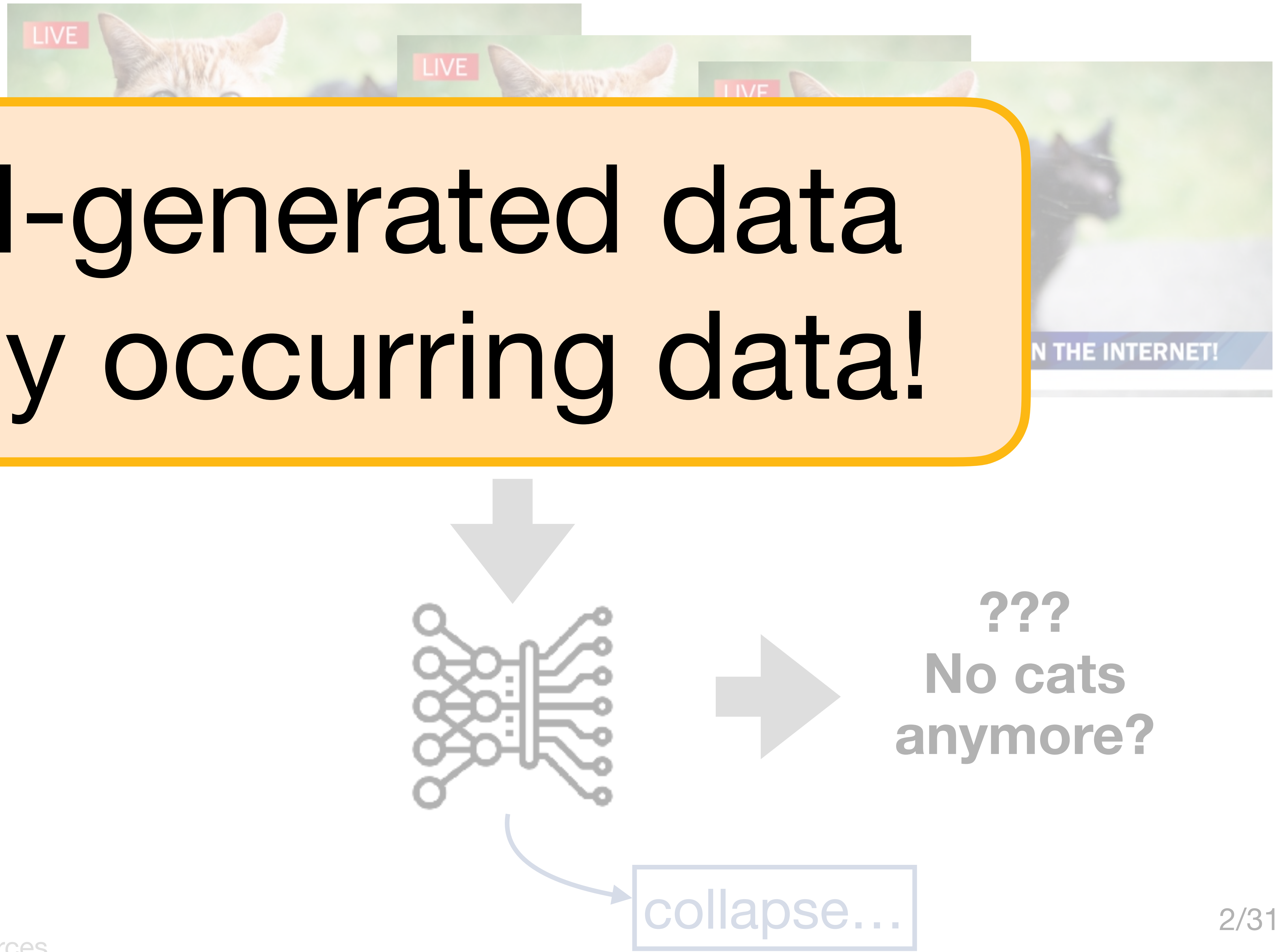
Challenges in AI Safety

Misuse of AI-generated content

Data Pollution

Plagiarism

Tons of AI-generate data over the internet



???
No cats anymore?

collapse…

We must distinguish AI-generated data from authentic, naturally occurring data!

[Source] Netskope.com and public sources.

Identify AI-generated Text

Possible solutions?

Identify AI-generated Text

Possible solutions?

- By observation:

Identify AI-generated Text

Possible solutions?

“*Here’s the revised version of your…*”, “*Best regards,[Your Name]*” :-D

Identify AI-generated Text

Possible solutions?

- Metadata <—easy to remove

Metadata

File name: Dataset
Author: GPT
Location: Ithaca
Created: Jan 01, 2025

Identify AI-generated Text

Possible solutions?

- Giant database to store all AI-generated content <—storage? privacy?

Identify AI-generated Text

Possible solutions?

- Discriminator models: GPTZero DetectGPT Copyleaks pangramlabs …

Identify AI-generated Text

Possible solutions?

<—high prob of falsely alarming human-written text

Identify AI-generated Text

Possible solutions?

- Watermarking: inserting a signal into LLM predicted tokens

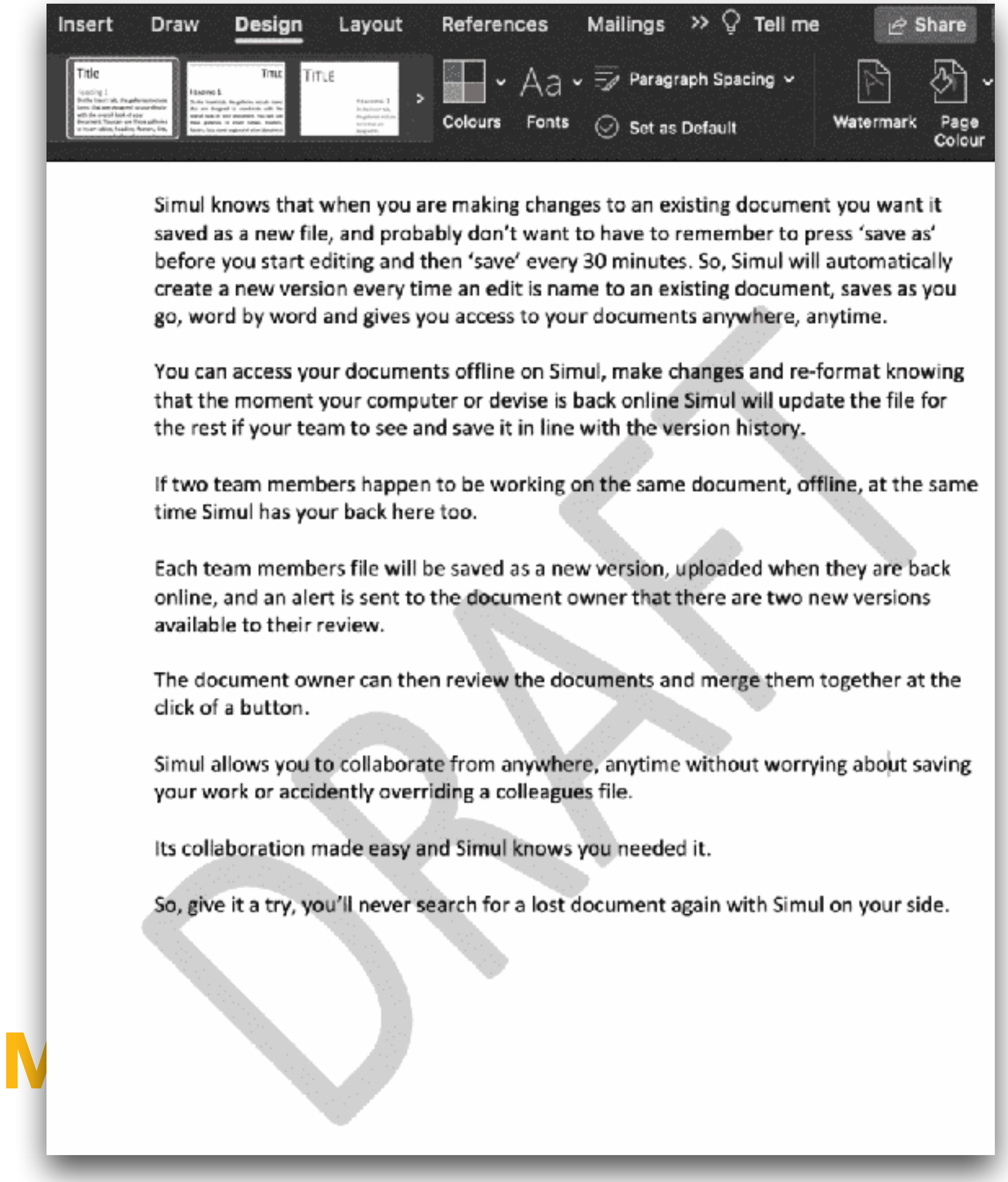
Identify AI-generated Text

Possible solutions?

- **Watermarking: inserting a signal into LLM predicted tokens**

Identify AI-generated Text

Possible solutions?



Simul knows that when you are making changes to an existing document you want it saved as a new file, and probably don't want to have to remember to press 'save as' before you start editing and then 'save' every 30 minutes. So, Simul will automatically create a new version every time an edit is name to an existing document, saves as you go, word by word and gives you access to your documents anywhere, anytime.

You can access your documents offline on Simul, make changes and re-format knowing that the moment your computer or devise is back online Simul will update the file for the rest if your team to see and save it in line with the version history.

If two team members happen to be working on the same document, offline, at the same time Simul has your back here too.

Each team members file will be saved as a new version, uploaded when they are back online, and an alert is sent to the document owner that there are two new versions available to their review.

The document owner can then review the documents and merge them together at the click of a button.

Simul allows you to collaborate from anywhere, anytime without worrying about saving your work or accidently overriding a colleagues file.

Its collaboration made easy and Simul knows you needed it.

So, give it a try, you'll never search for a lost document again with Simul on your side.

- **Watermarking: inserting a signal into LLM**

Identify AI-generated Text

Possible solutions?

- **Watermarking: inserting a signal into LLM predicted tokens**

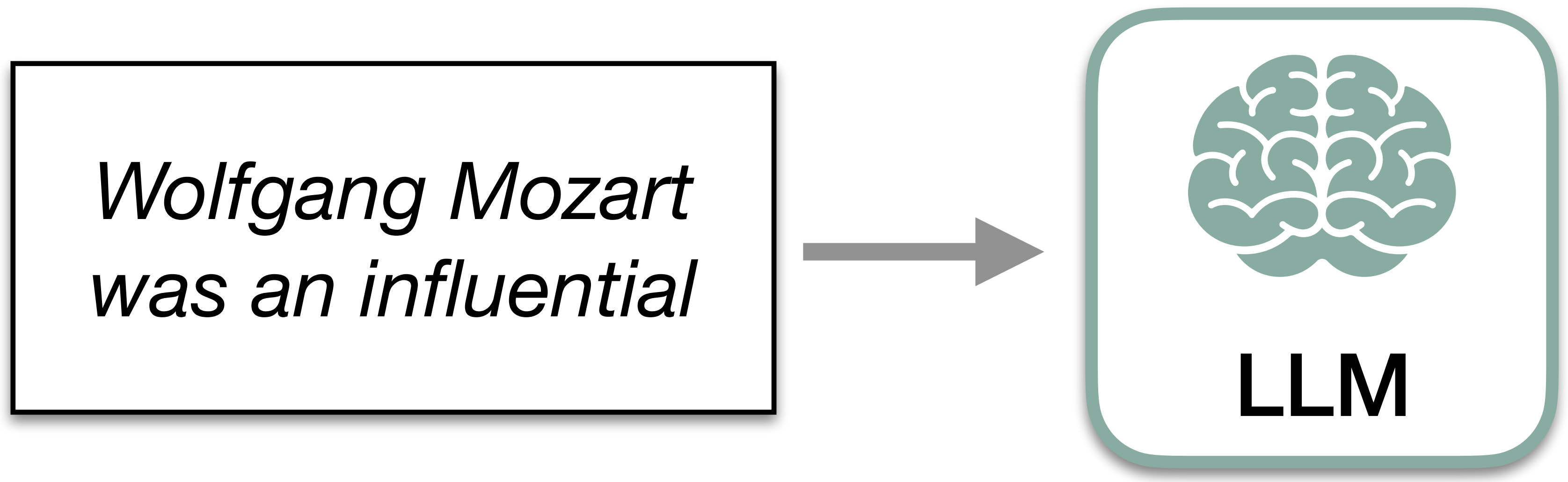
Identify AI-generated Text

Possible solutions?

- **Watermarking: inserting a signal into LLM predicted tokens**

A Framework for LLM Watermark Generation

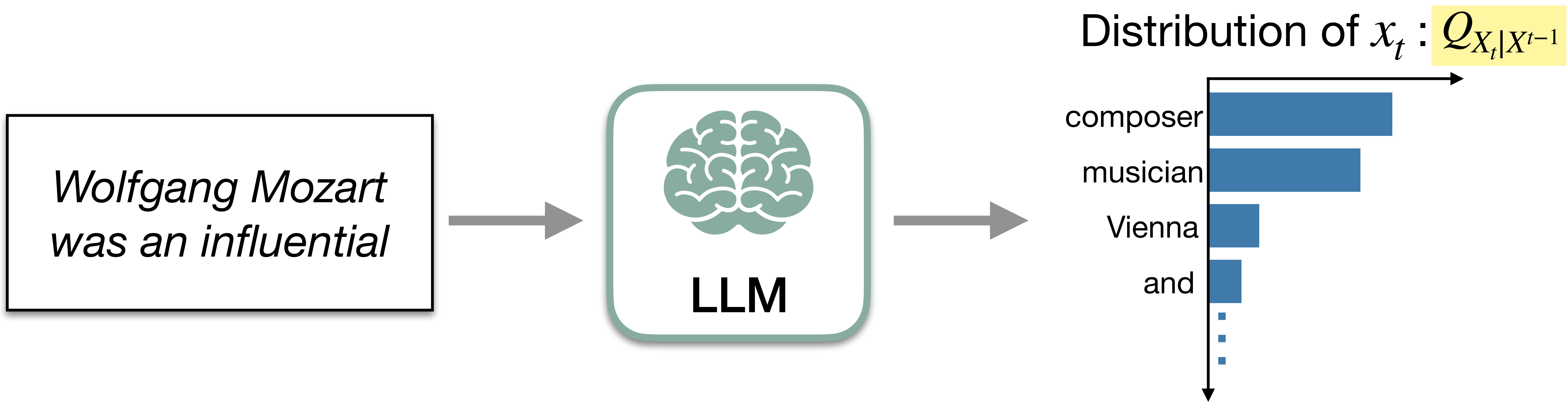
A Framework for LLM Watermark Generation



Wolfgang Mozart was an influential

LLM

A Framework for LLM Watermark Generation

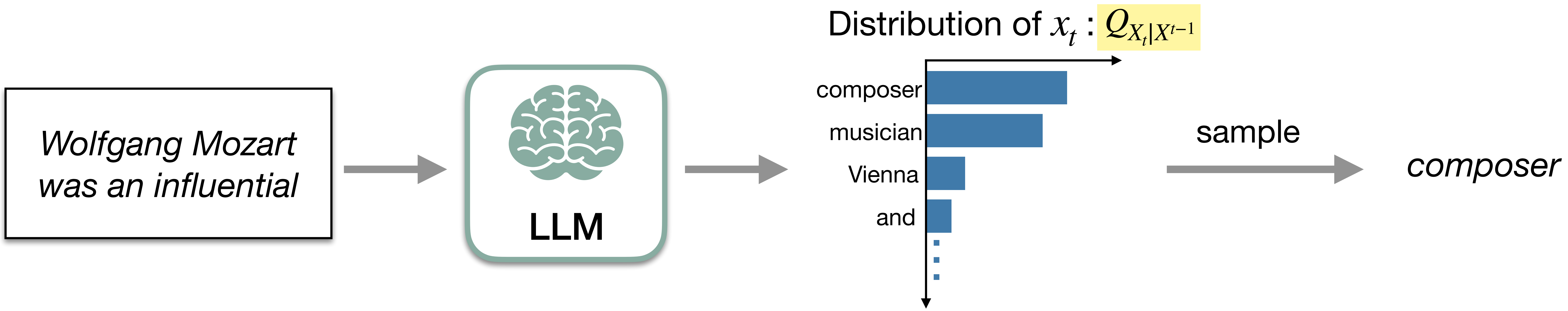


Wolfgang Mozart was an influential

LLM

Distribution of x_t : $Q_{X_t|X^{t-1}}$

A Framework for LLM Watermark Generation



Wolfgang Mozart was an influential

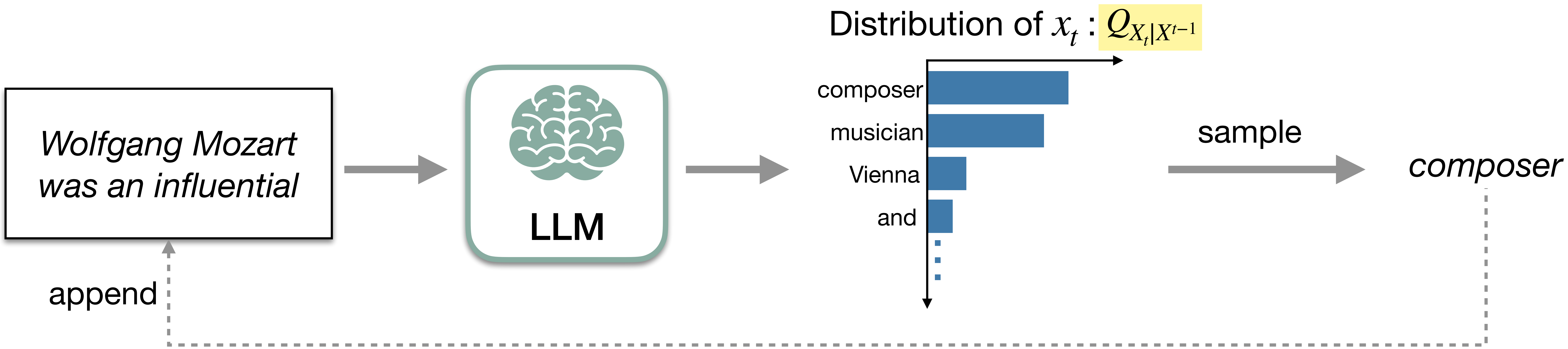
LLM

Distribution of x_t : $Q_{X_t|X^{t-1}}$

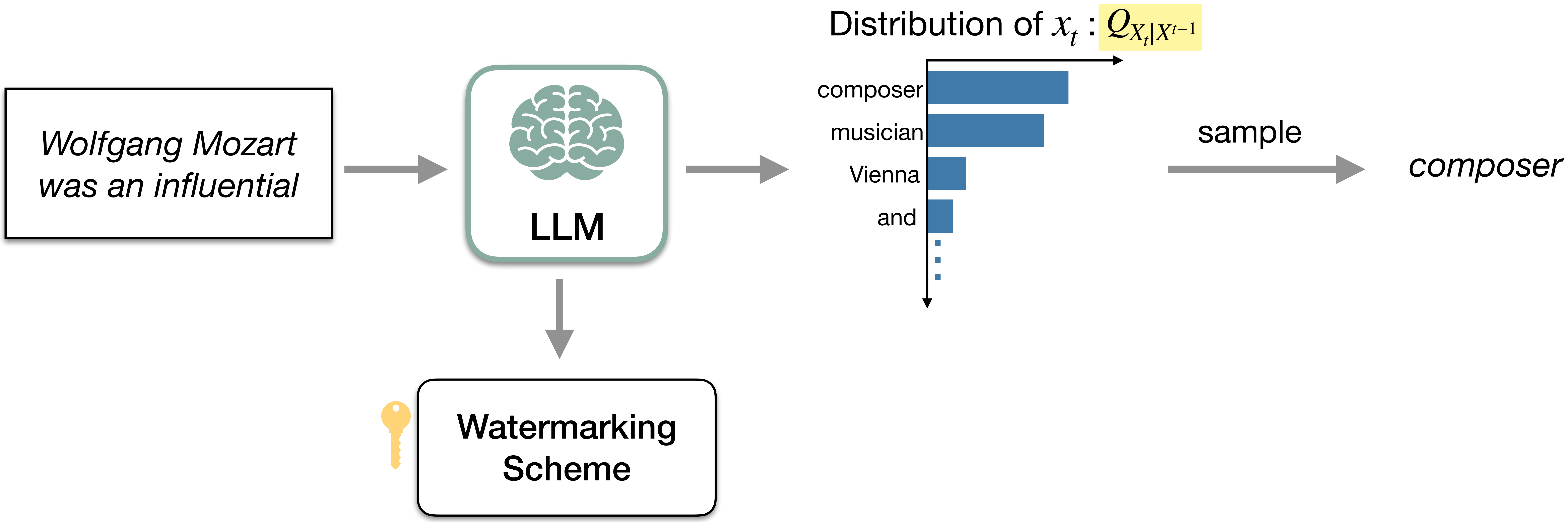
sample

composer

A Framework for LLM Watermark Generation


Wolfgang Mozart was an influential
LLM
Distribution of x_t : $Q_{X_t|X^{t-1}}$
composer
musician
Vienna
and
sample
composer
append

A Framework for LLM Watermark Generation



Wolfgang Mozart was an influential

LLM

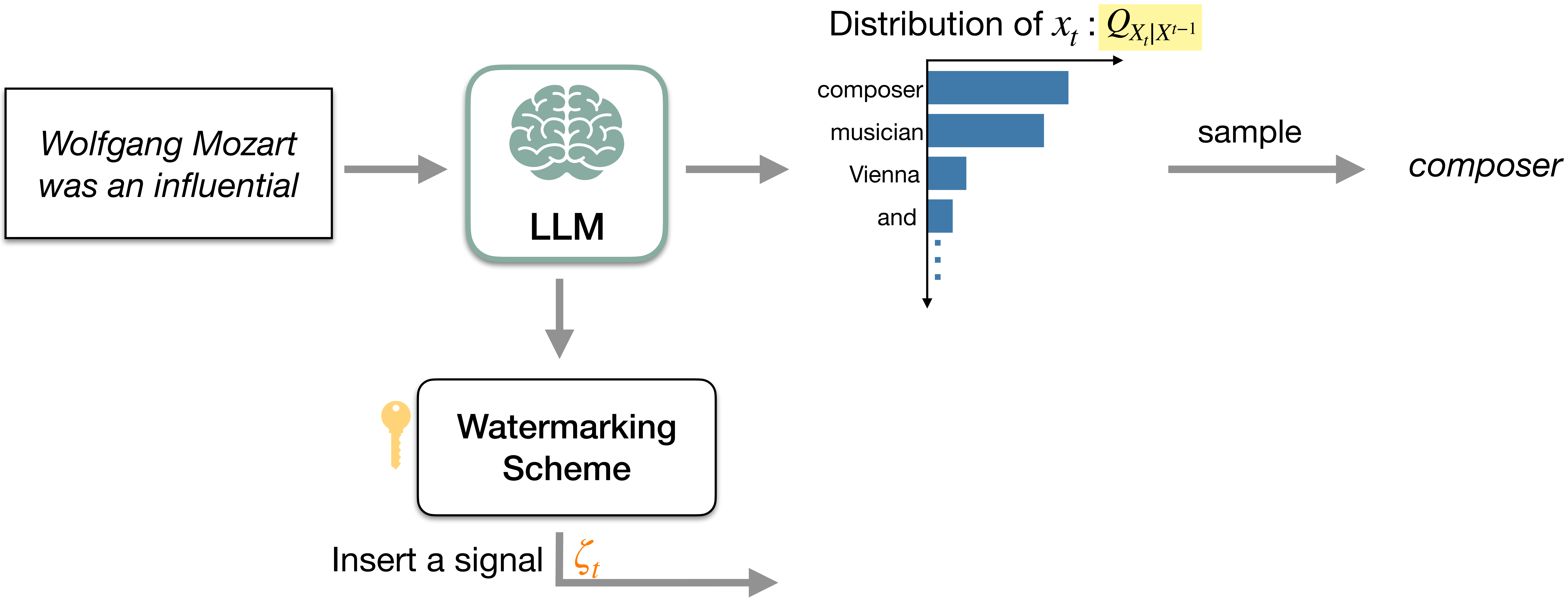
Distribution of x_t : $Q_{X_t|X^{t-1}}$

sample

composer

Watermarking Scheme

A Framework for LLM Watermark Generation



Wolfgang Mozart was an influential

LLM

Distribution of x_t : $Q_{X_t|X^{t-1}}$

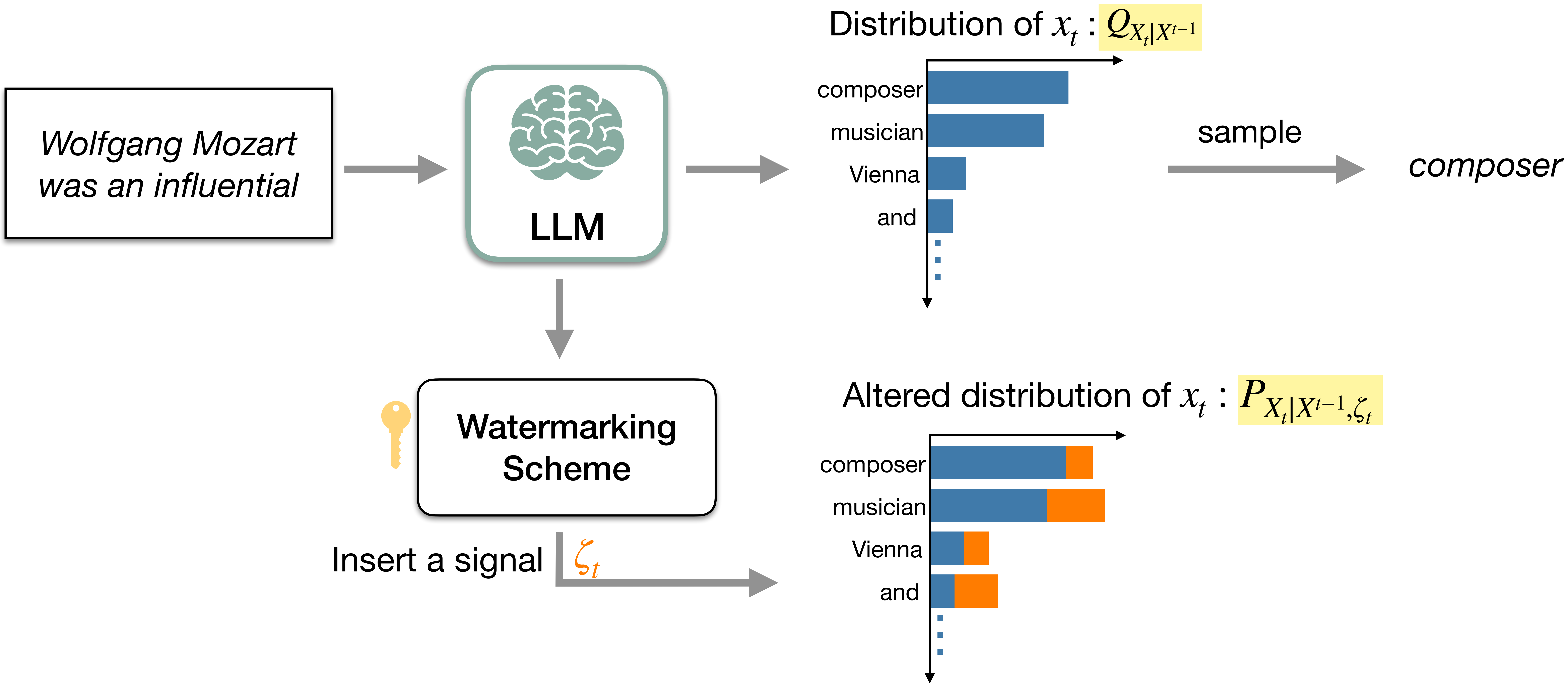
sample

composer

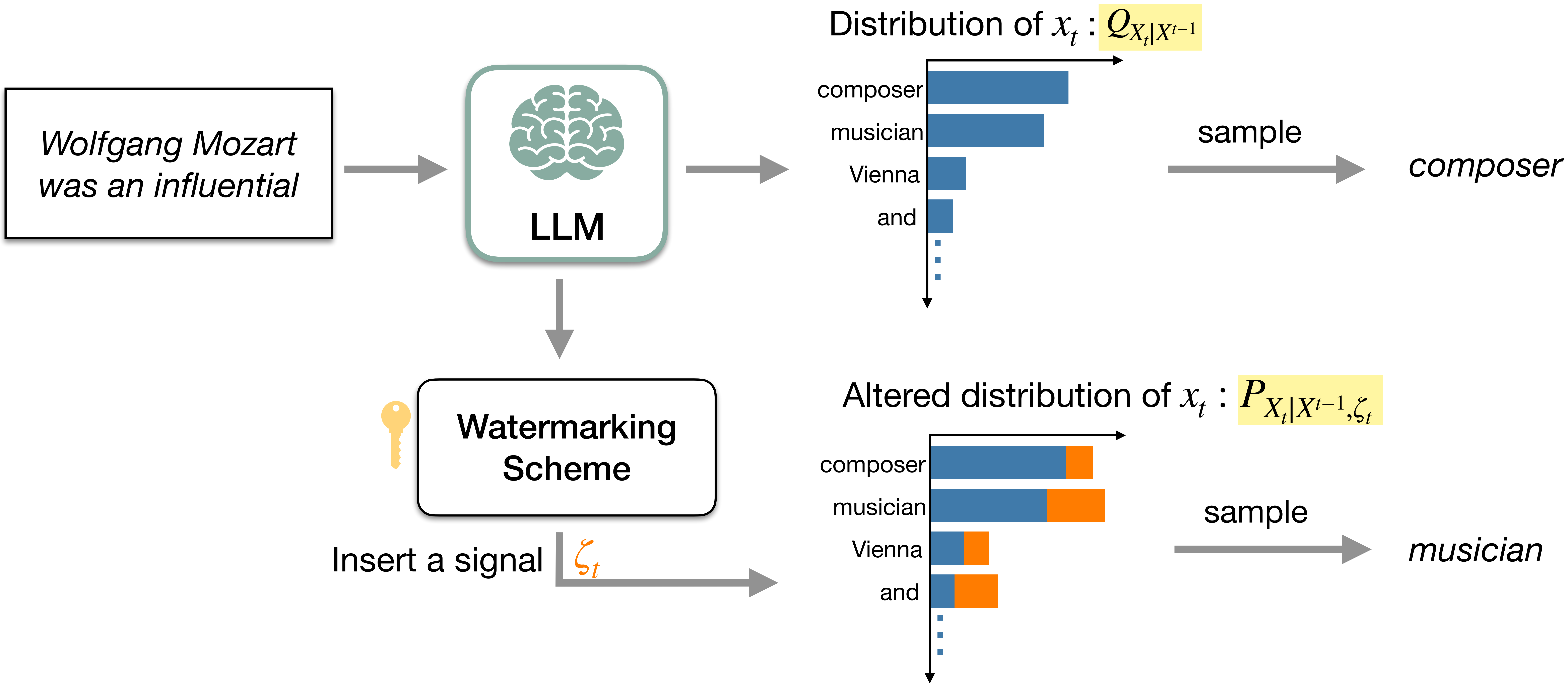
Watermarking Scheme

Insert a signal ζ_t

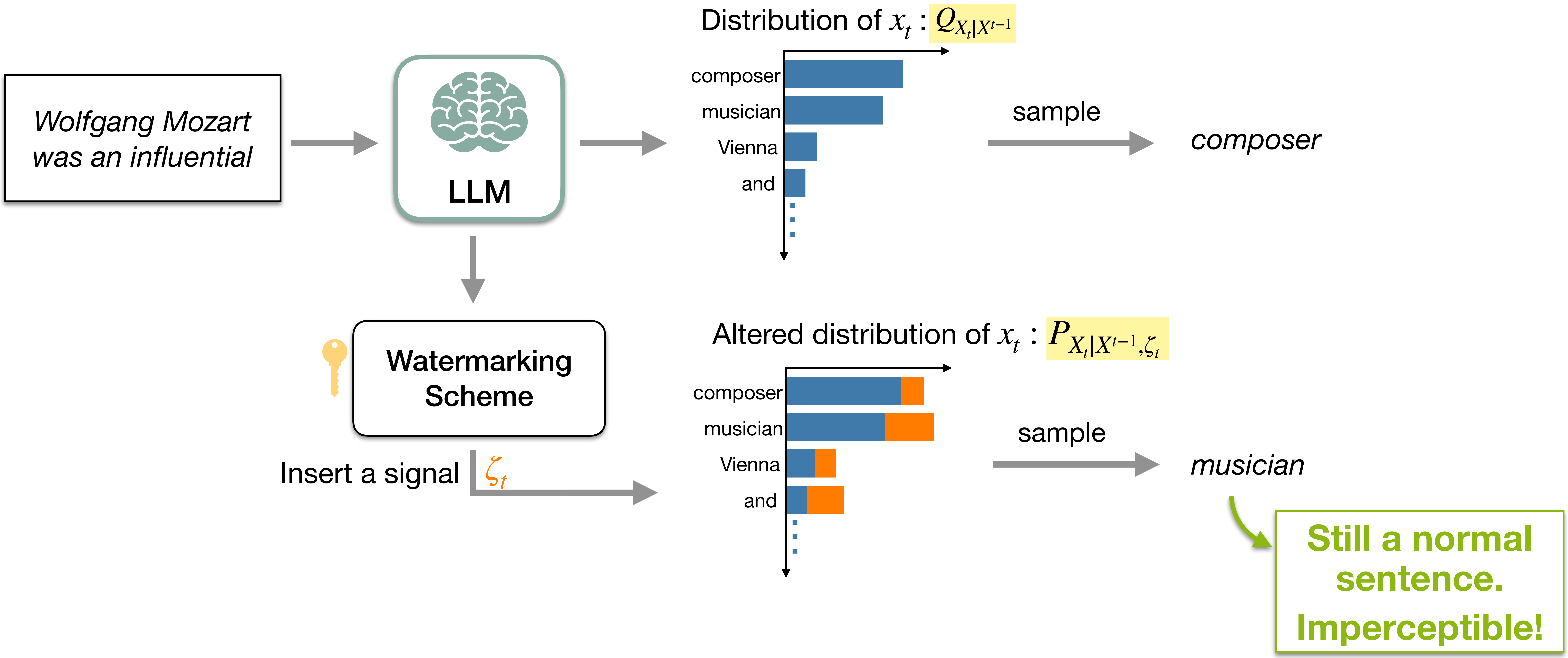
A Framework for LLM Watermark Generation


Wolfgang Mozart was an influential
LLM
Distribution of x_t : $Q_{X_t|X^{t-1}}$
composer
musician
Vienna
and
sample
composer
Watermarking Scheme
Insert a signal ζ_t
Altered distribution of x_t : $P_{X_t|X^{t-1},\zeta_t}$
composer
musician
Vienna
and

A Framework for LLM Watermark Generation

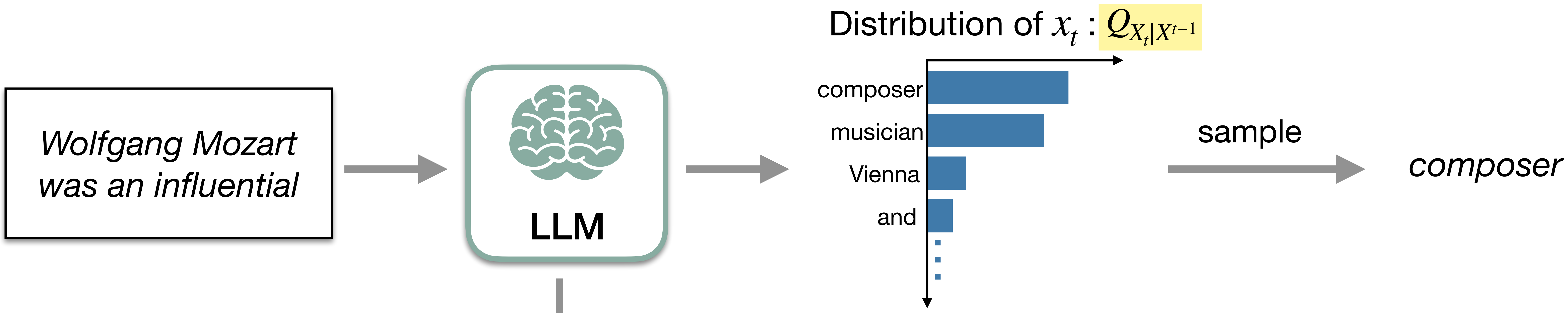

Wolfgang Mozart was an influential
LLM
Distribution of x_t : $Q_{X_t|X^{t-1}}$
composer
musician
Vienna
and
sample
composer
Watermarking Scheme
Insert a signal ζ_t
Altered distribution of x_t : $P_{X_t|X^{t-1},\zeta_t}$
composer
musician
Vienna
and
sample
musician

A Framework for LLM Watermark Generation

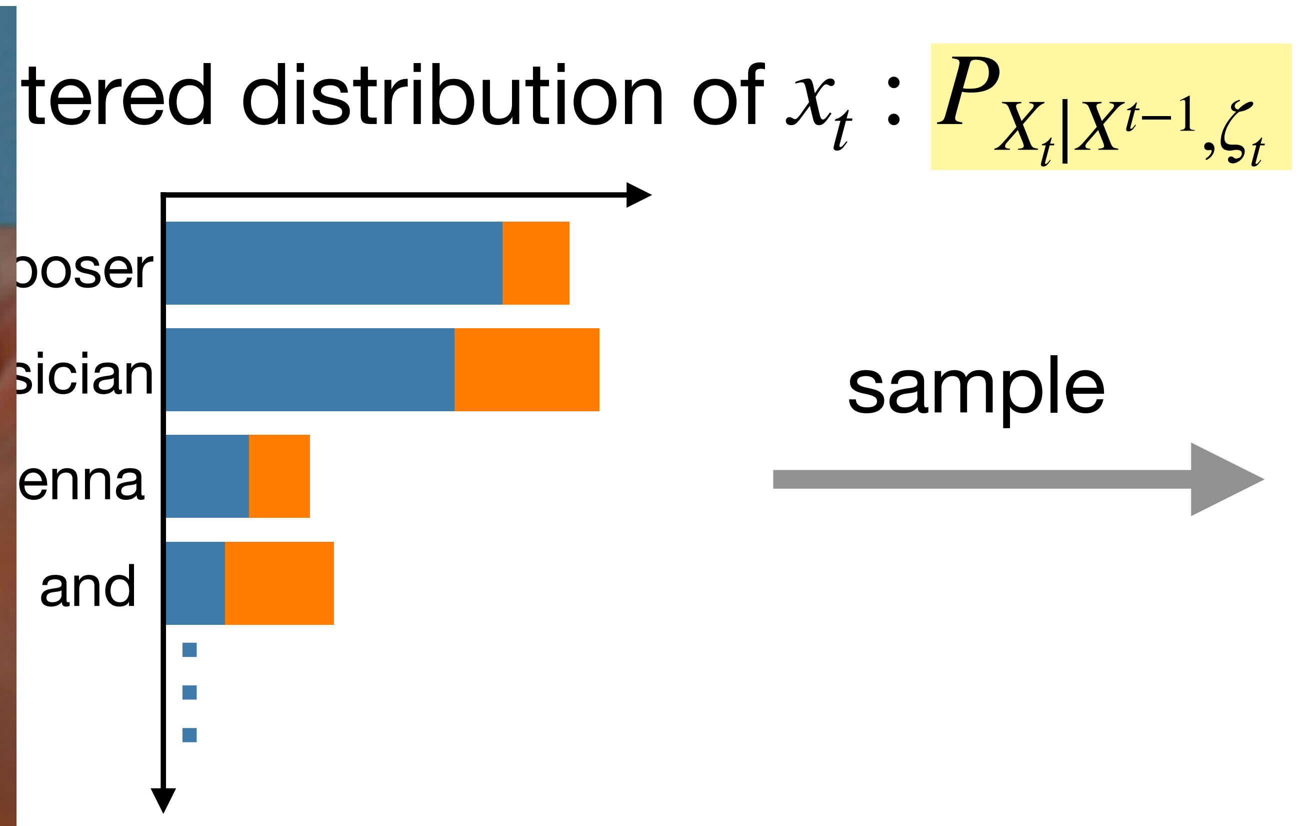

Wolfgang Mozart was an influential
LLM
Distribution of x_t : $Q_{X_t|X^{t-1}}$
composer
musician
Vienna
and
sample
composer
Watermarking Scheme
Insert a signal ζ_t
Altered distribution of x_t : $P_{X_t|X^{t-1},\zeta_t}$
composer
musician
Vienna
and
sample
musician

Still a normal sentence. Imperceptible!

A Framework for LLM Watermark Generation

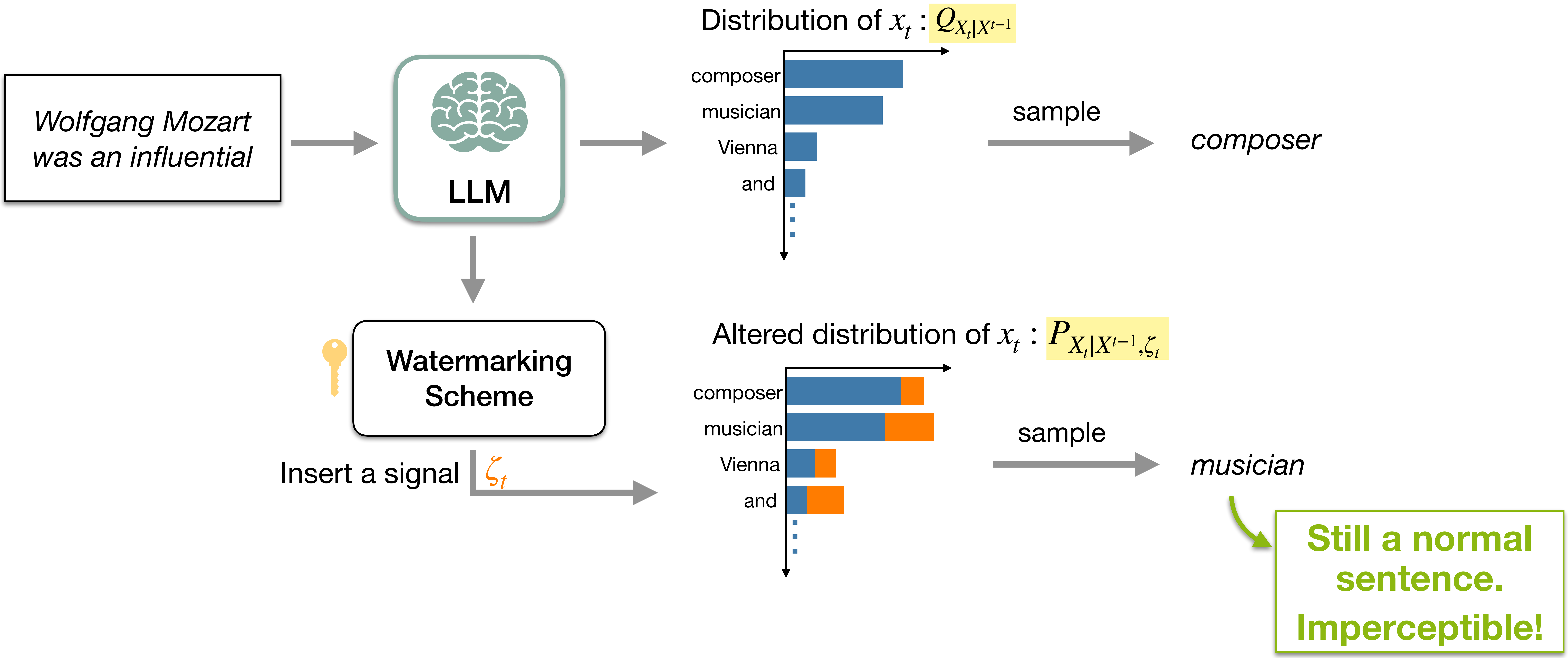

Wolfgang Mozart was an influential
LLM
Distribution of x_t : $Q_{X_t|X^{t-1}}$
composer
musician
Vienna
and
sample
composer

Like invisible Ink (Steganography)



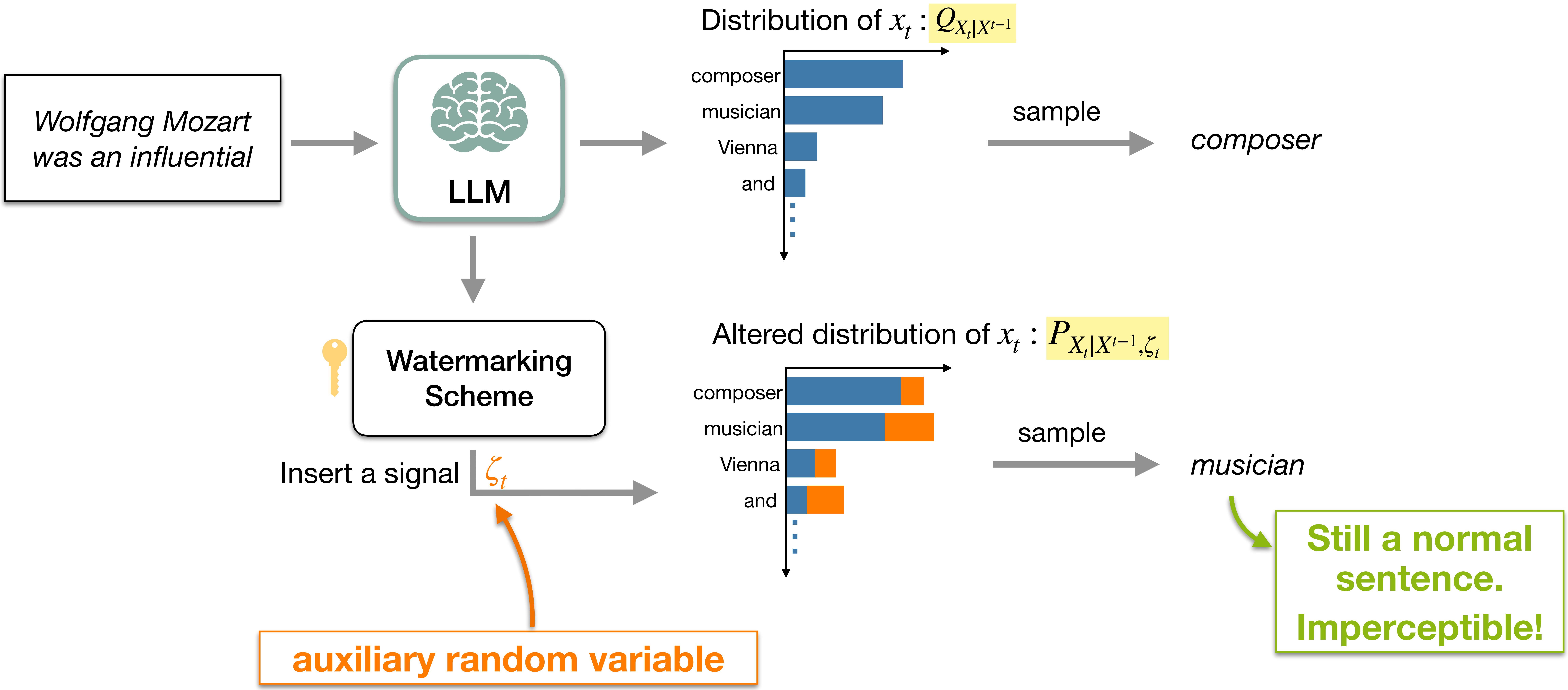
Still a normal sentence. Imperceptible!

A Framework for LLM Watermark Generation


Wolfgang Mozart was an influential
LLM
Distribution of x_t : $Q_{X_t|X^{t-1}}$
composer
musician
Vienna
and
sample
composer
Watermarking Scheme
Insert a signal ζ_t
Altered distribution of x_t : $P_{X_t|X^{t-1},\zeta_t}$
composer
musician
Vienna
and
sample
musician

Still a normal sentence. Imperceptible!

A Framework for LLM Watermark Generation



Wolfgang Mozart was an influential

LLM

Distribution of x_t : $Q_{X_t|X^{t-1}}$

sample

composer

Watermarking Scheme

Insert a signal ζ_t

auxiliary random variable

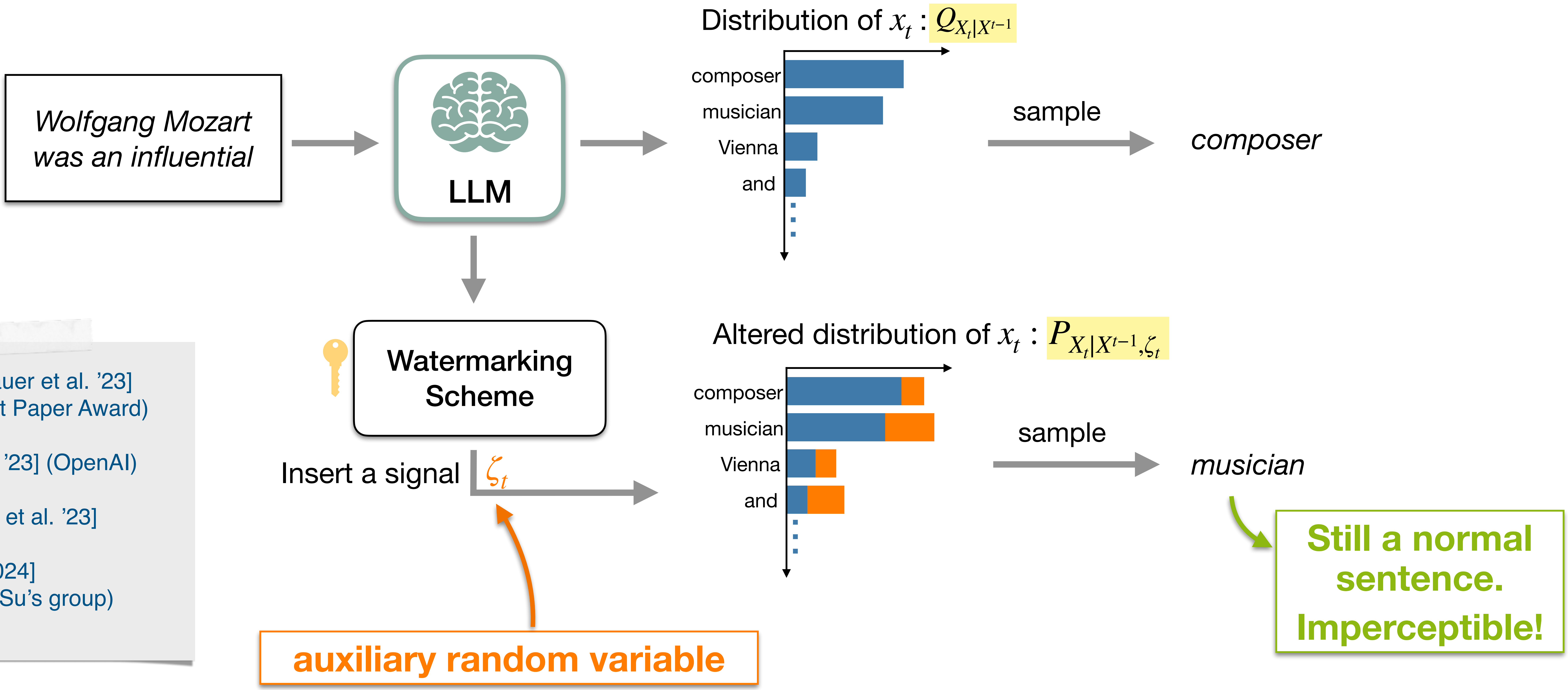
Altered distribution of x_t : $P_{X_t|X^{t-1},\zeta_t}$

sample

musician

Still a normal sentence. Imperceptible!

A Framework for LLM Watermark Generation



Wolfgang Mozart was an influential

Distribution of x_t : $Q_{X_t|X^{t-1}}$

Watermarking Scheme

Insert a signal ζ_t

Altered distribution of x_t : $P_{X_t|X^{t-1},\zeta_t}$

auxiliary random variable

Still a normal sentence. Imperceptible!

[Kirchenbauer et al. ’23] (ICML Best Paper Award)

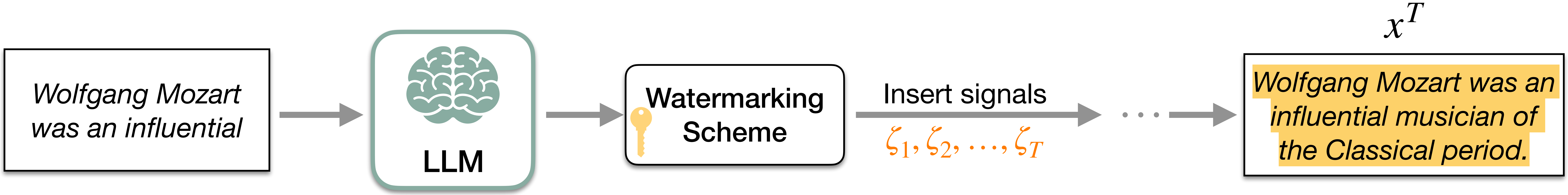
[Aaronson ’23] (OpenAI)

[Kuditipudi et al. ’23]

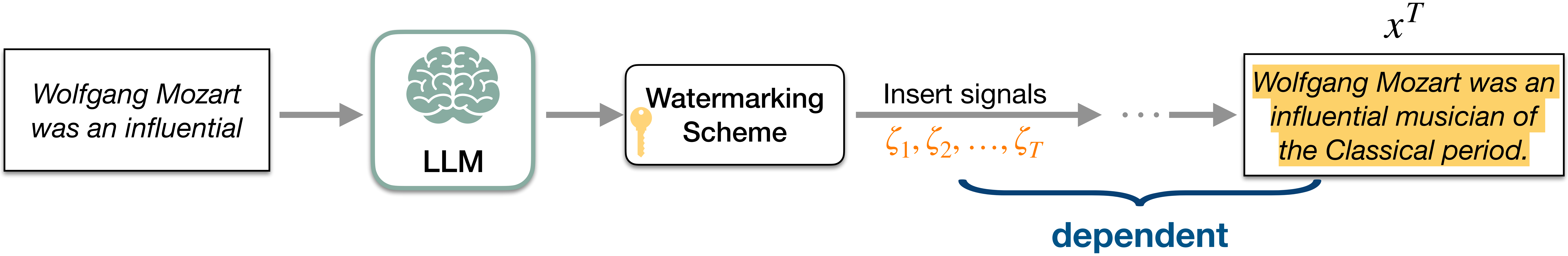
[Li et al. 2024] (by Weijie Su’s group)

…

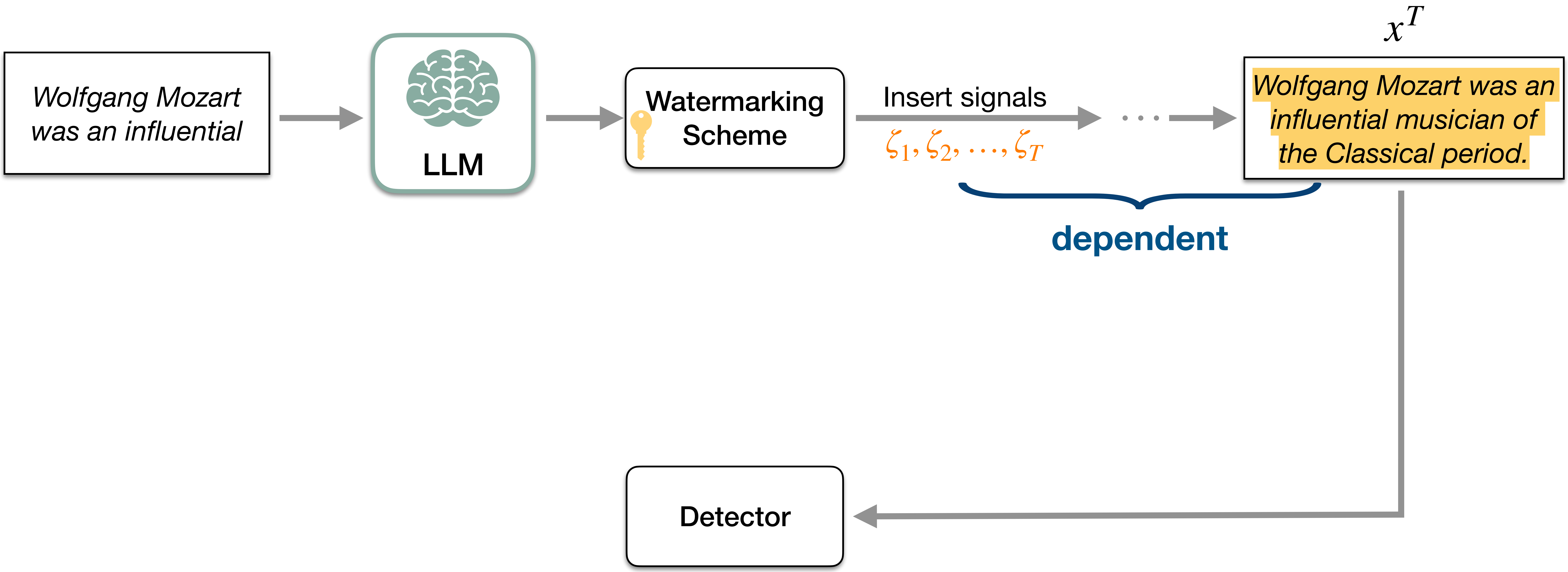
Hypothesis Testing for LLM Watermark Detection


Wolfgang Mozart was an influential
LLM
Watermarking Scheme
Insert signals
$\zeta_1, \zeta_2, \ldots, \zeta_T$
$\cdots$
x^T
Wolfgang Mozart was an influential musician of the Classical period.

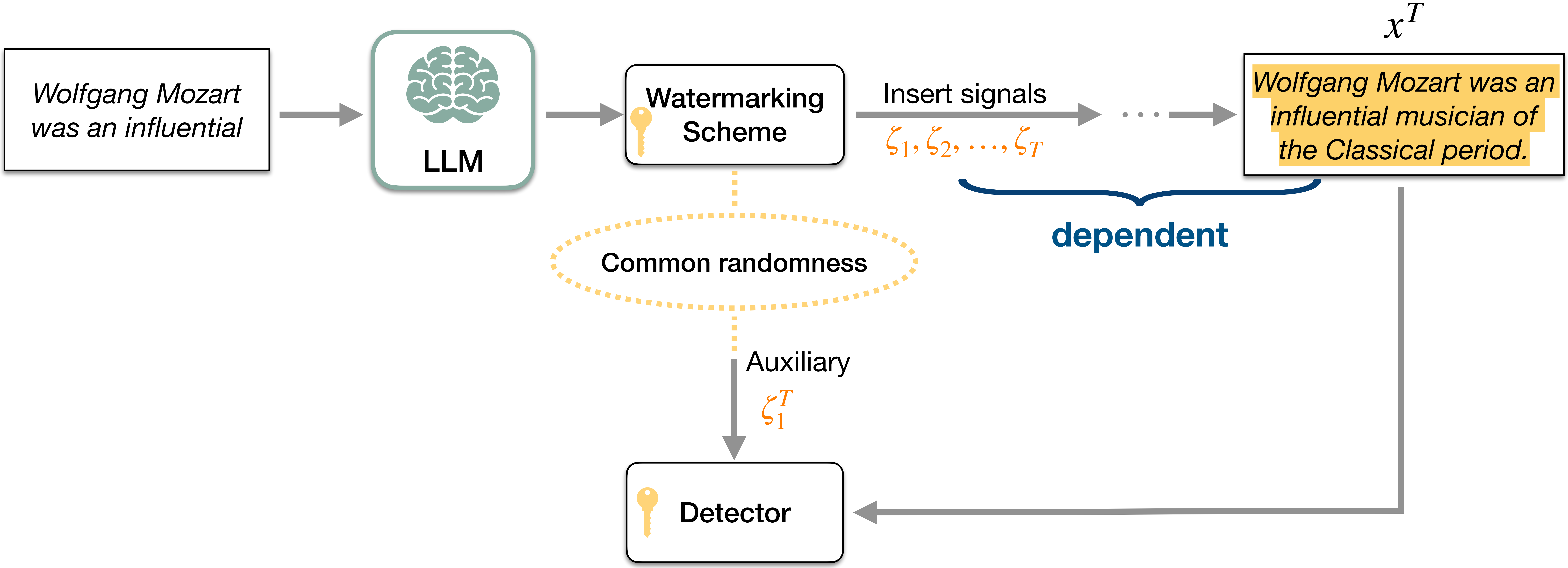
Hypothesis Testing for LLM Watermark Detection


Wolfgang Mozart was an influential
LLM
Watermarking Scheme
Insert signals
$\zeta_1, \zeta_2, \ldots, \zeta_T$
$\cdots$
x^T
Wolfgang Mozart was an influential musician of the Classical period.
dependent

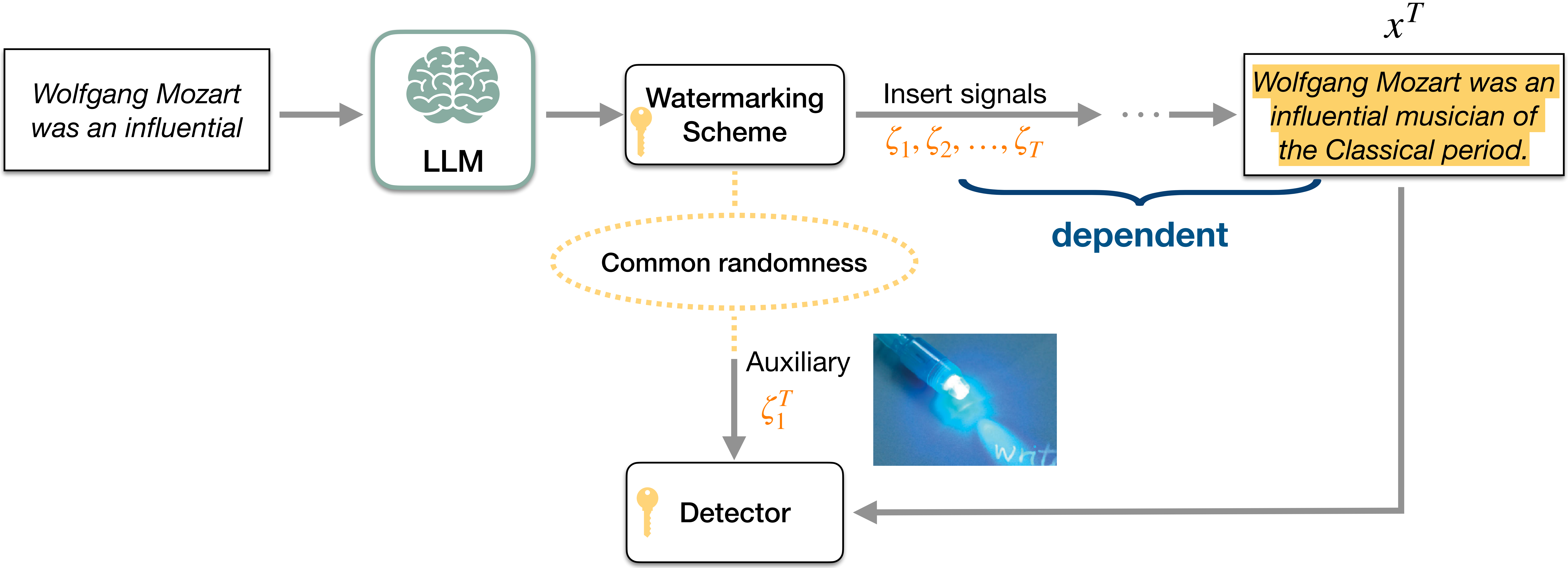
Hypothesis Testing for LLM Watermark Detection


Wolfgang Mozart was an influential
LLM
Watermarking Scheme
Insert signals
$\zeta_1, \zeta_2, \ldots, \zeta_T$
$\cdots$
x^T
Wolfgang Mozart was an influential musician of the Classical period.
dependent
Detector

Hypothesis Testing for LLM Watermark Detection


Wolfgang Mozart was an influential
LLM
Watermarking Scheme
Insert signals
$\zeta_1, \zeta_2, \ldots, \zeta_T$
$\cdots$
x^T
Wolfgang Mozart was an influential musician of the Classical period.
dependent
Common randomness
Auxiliary
ζ_1^T
Detector

Hypothesis Testing for LLM Watermark Detection



Wolfgang Mozart was an influential

LLM

Watermarking Scheme

Insert signals

$\zeta_1, \zeta_2, \ldots, \zeta_T$

$\cdots$

x^T

Wolfgang Mozart was an influential musician of the Classical period.

dependent

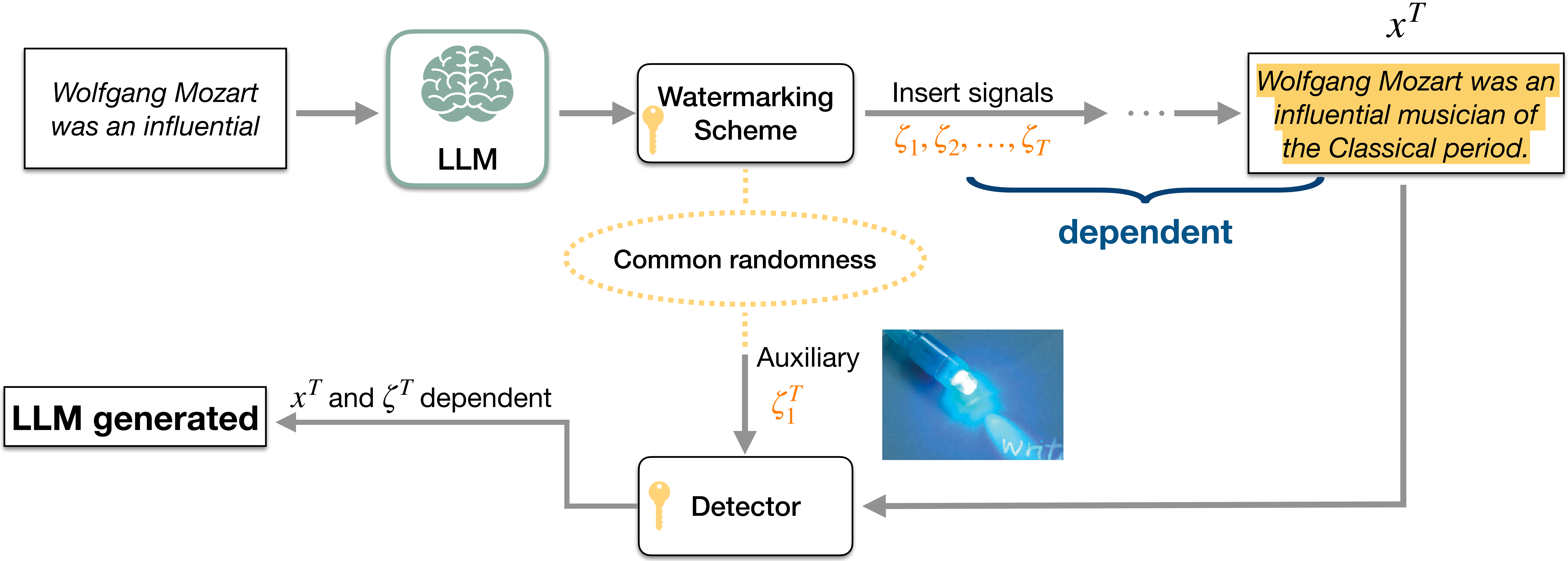
Common randomness

Auxiliary

ζ_1^T

Detector

Hypothesis Testing for LLM Watermark Detection



Wolfgang Mozart was an influential

LLM

Watermarking Scheme

Insert signals

$\zeta_1, \zeta_2, \ldots, \zeta_T$

$\cdots$

x^T

Wolfgang Mozart was an influential musician of the Classical period.

dependent

Common randomness

Auxiliary

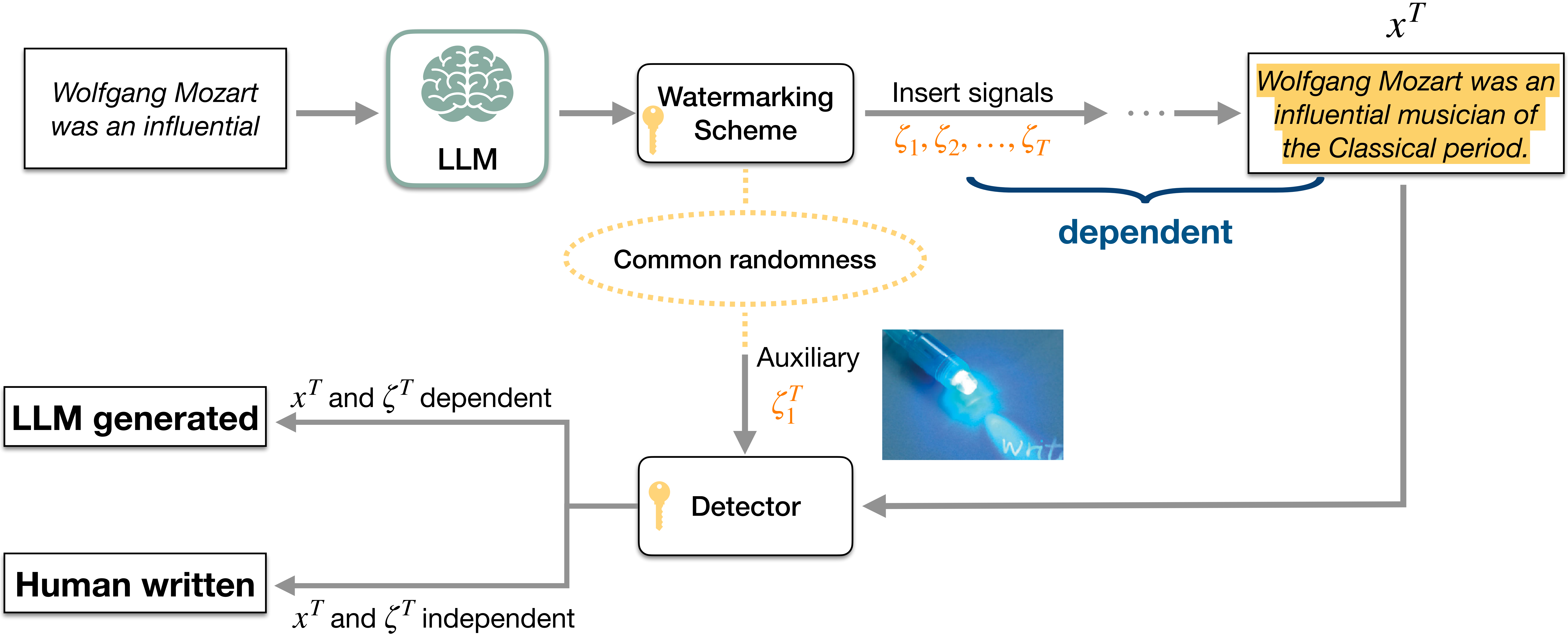
ζ_1^T

Detector

x^T and ζ^T dependent

LLM generated

Hypothesis Testing for LLM Watermark Detection



Wolfgang Mozart was an influential

LLM

Watermarking Scheme

Insert signals

$\zeta_1, \zeta_2, \ldots, \zeta_T$

$\cdots$

x^T

Wolfgang Mozart was an influential musician of the Classical period.

dependent

Common randomness

Auxiliary

ζ_1^T

Detector

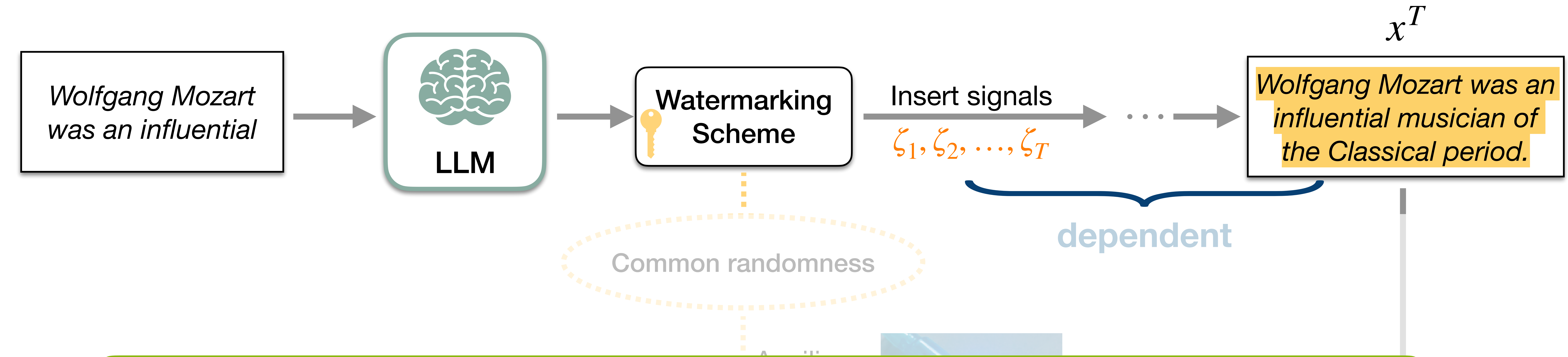
LLM generated

x^T and ζ^T dependent

Human written

x^T and ζ^T independent

Hypothesis Testing for LLM Watermark Detection


Wolfgang Mozart was an influential
LLM
Watermarking Scheme
Insert signals
$\zeta_1, \zeta_2, \ldots, \zeta_T$
$\cdots$
x^T
Wolfgang Mozart was an influential musician of the Classical period.
dependent
Common randomness

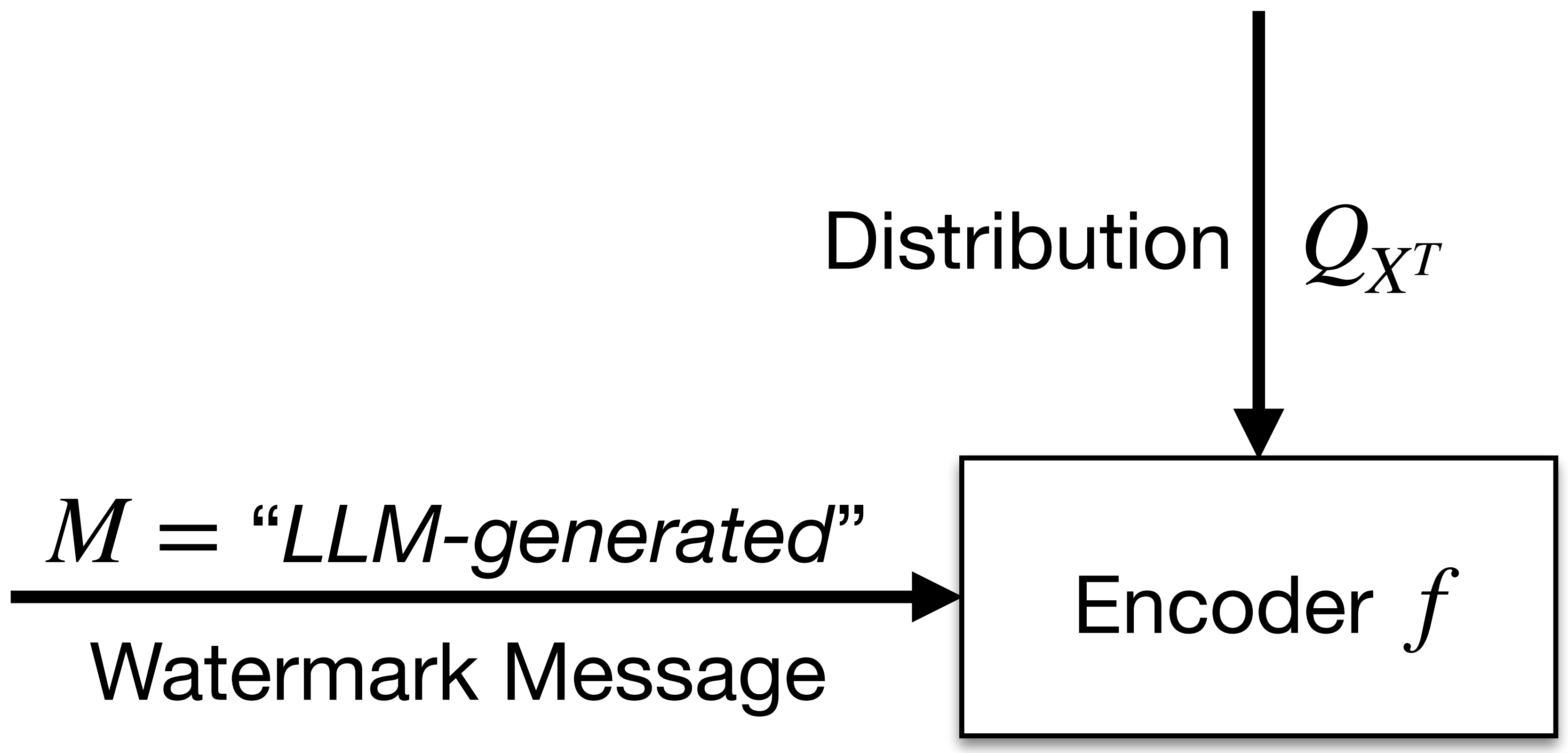
Watermark Detection $\Longrightarrow$ Hypothesis Testing:

H_0 : X^T is human written, i.e., $(X^T, \zeta^T) \sim Q_{X^T} \otimes P_{\zeta^T}$

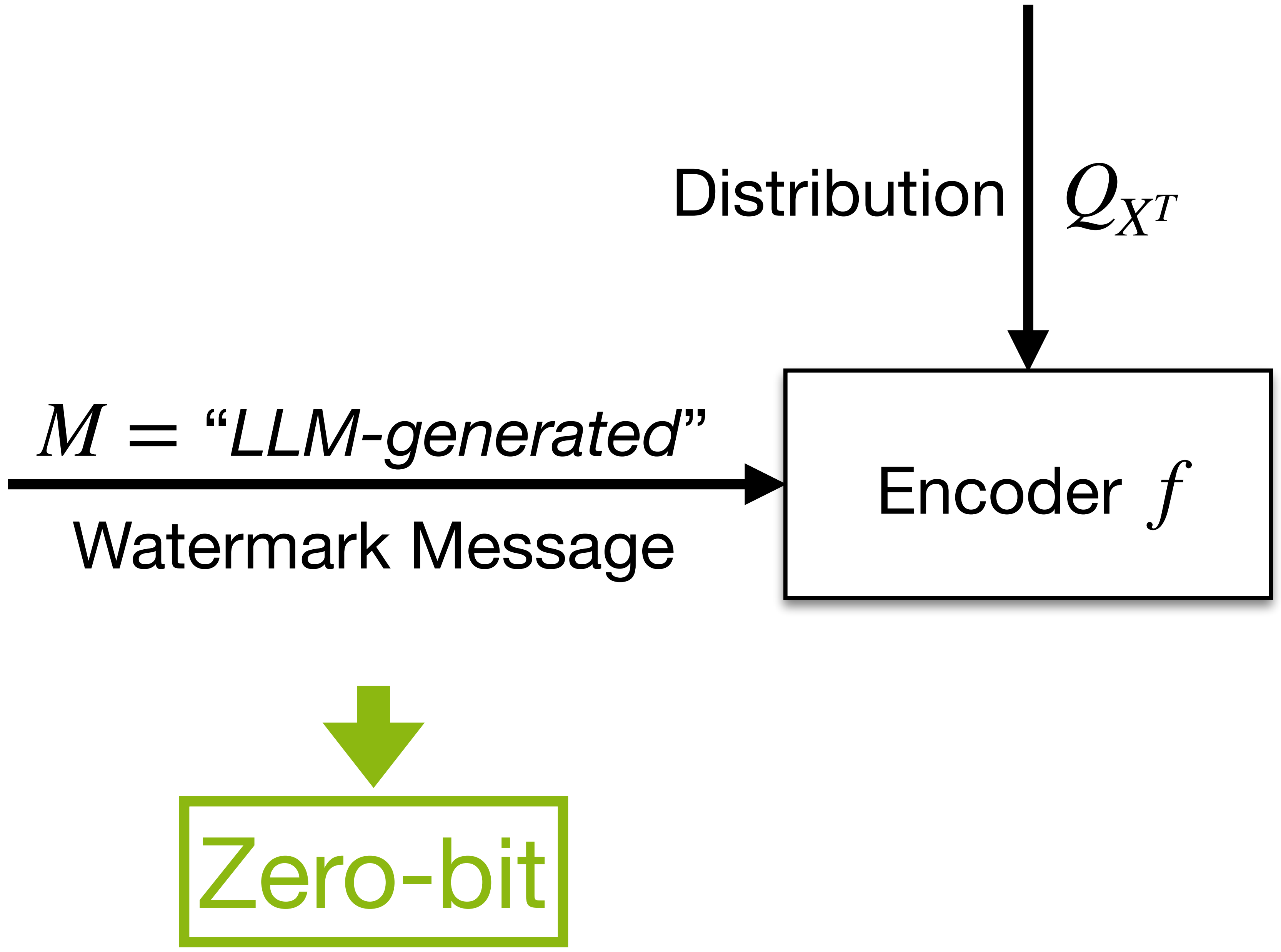
H_1 : X^T is LLM generated, i.e., $(X^T, \zeta^T) \sim P_{X^T, \zeta^T}$

Framework: Distributional Information Embedding with Side Information

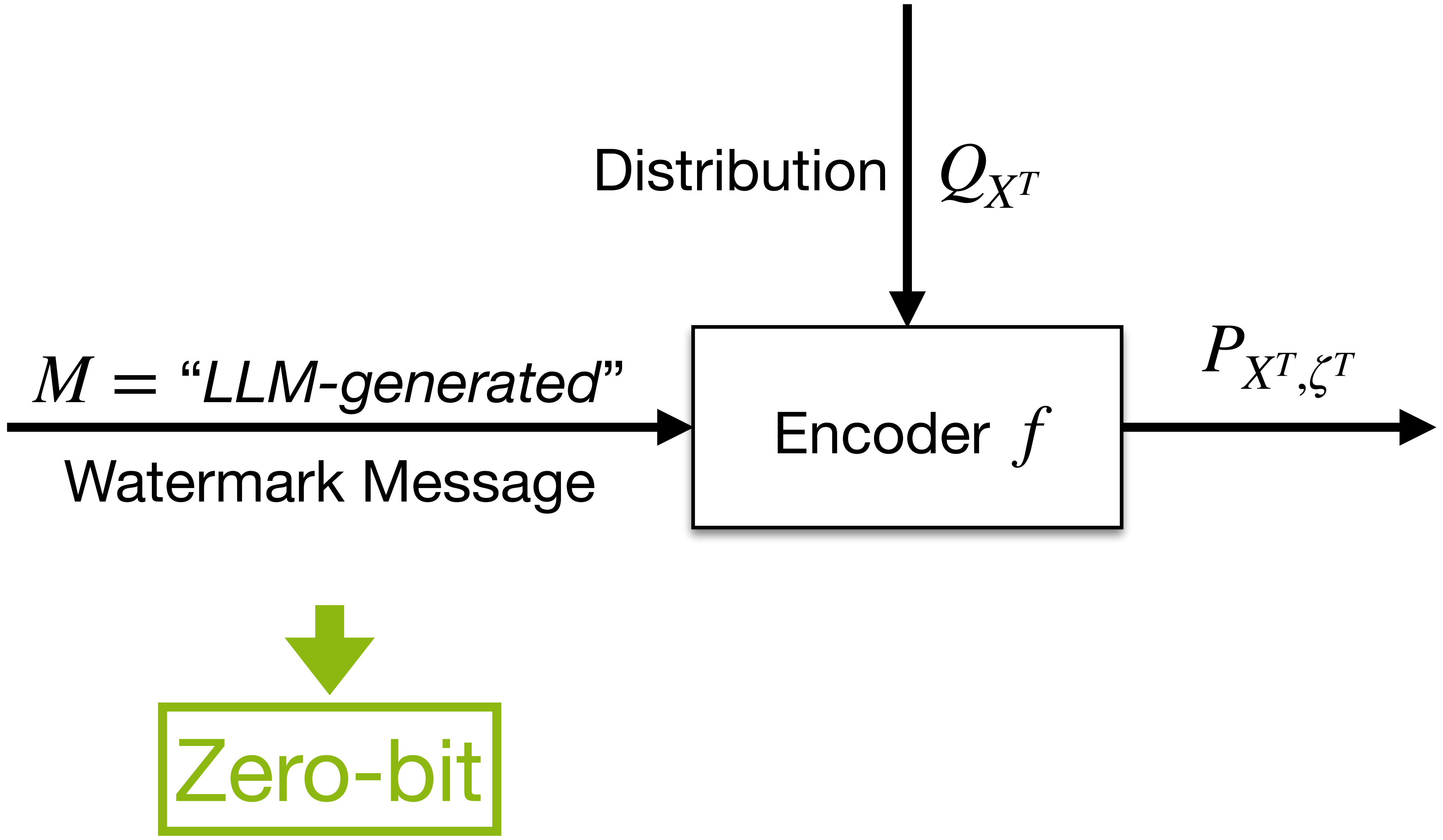
Framework: Distributional Information Embedding with Side Information



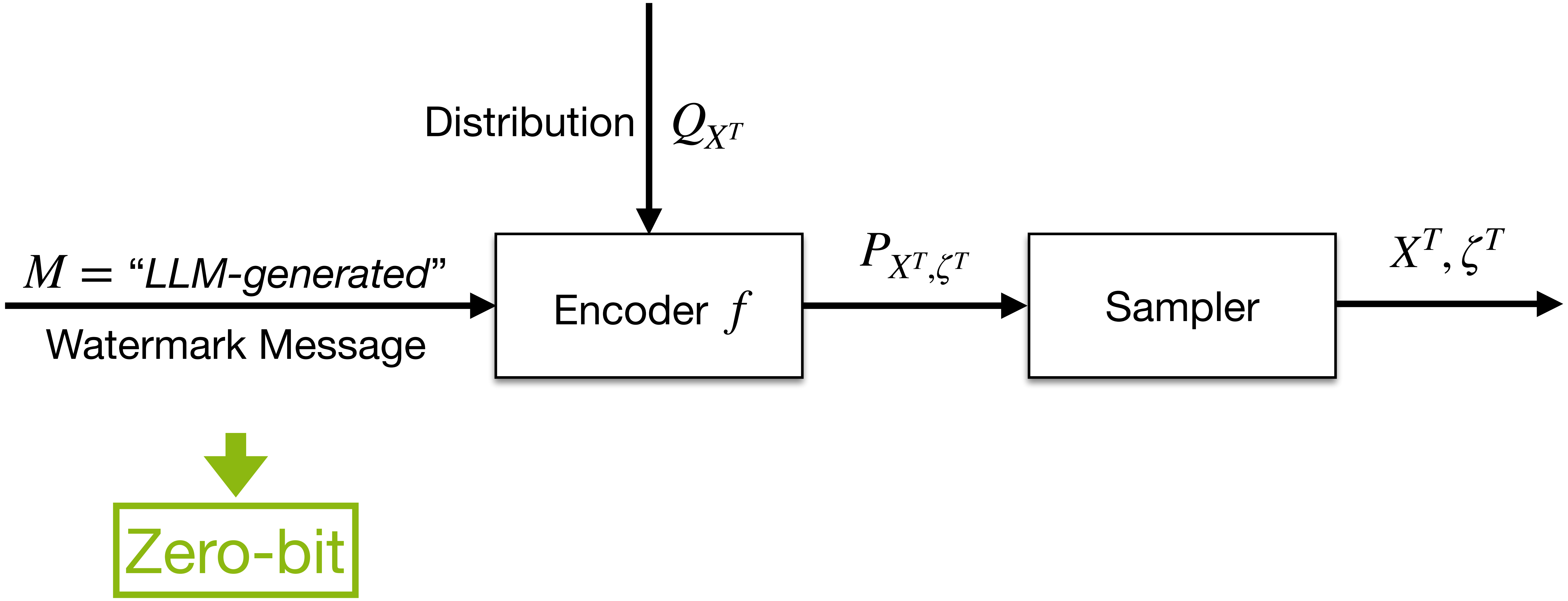
Framework: Distributional Information Embedding with Side Information



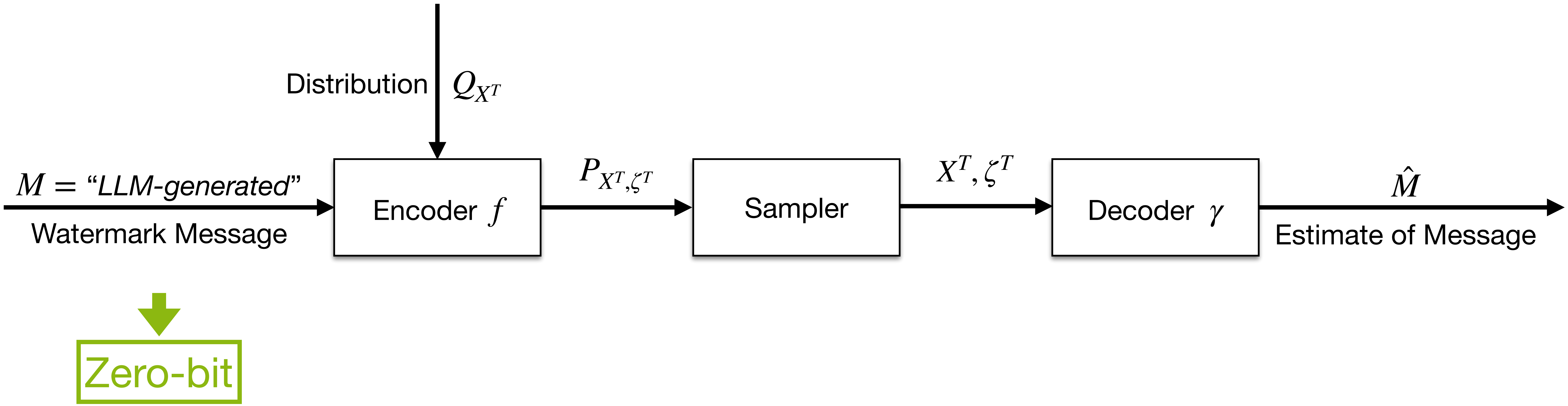
Framework: Distributional Information Embedding with Side Information



Framework: Distributional Information Embedding with Side Information



Framework: Distributional Information Embedding with Side Information



Distribution Q_{X^T}

$M =$ *“LLM-generated”*

Watermark Message

Encoder f

P_{X^T,ζ^T}

Sampler

X^T, ζ^T

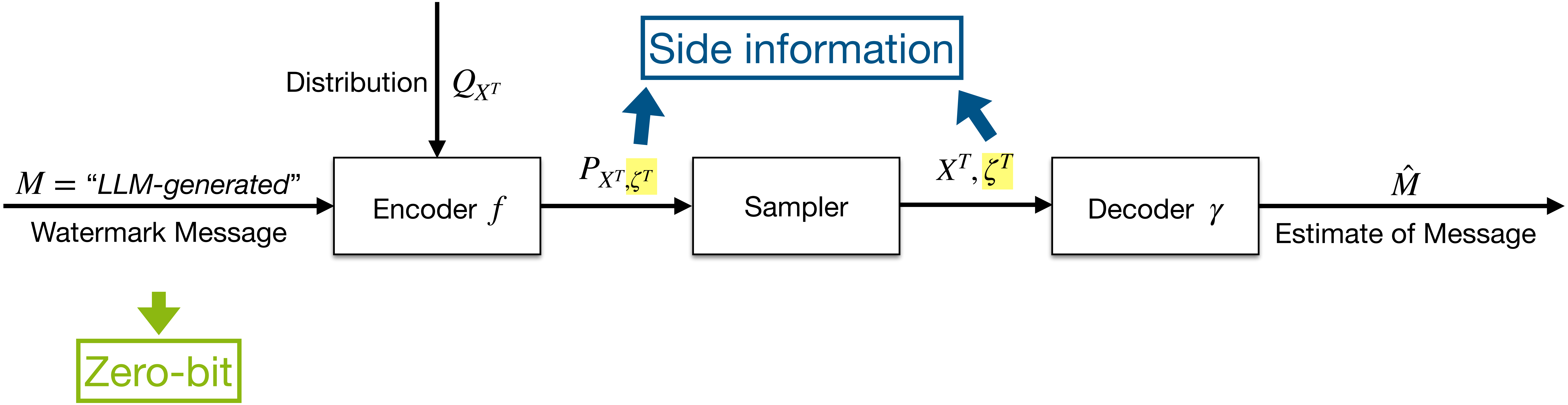
Decoder γ

$\hat{M}$

Estimate of Message

Zero-bit

Framework: Distributional Information Embedding with Side Information



LLM Watermark Detection Errors

Watermark Detection $\Longrightarrow$ Hypothesis Testing:

H_0 : X^T is human written, i.e., $(X^T, \zeta^T) \sim Q_{X^T} \otimes P_{\zeta^T}$

H_1 : X^T is LLM generated, i.e., $(X^T, \zeta^T) \sim P_{X^T, \zeta^T}$

LLM Watermark Detection Errors

Watermark Detection ⟹ Hypothesis Testing: Human/unwatermarked LLM

$\mathrm{H}_0 : X^T$ is human written, i.e., $(X^T, \zeta^T) \sim \boxed{Q_{X^T}} \otimes P_{\zeta^T}$

$\mathrm{H}_1 : X^T$ is LLM generated, i.e., $(X^T, \zeta^T) \sim P_{X^T, \zeta^T}$

LLM Watermark Detection Errors

Watermark Detection $\Longrightarrow$ Hypothesis Testing: Human/unwatermarked LLM

H_0 : X^T is human written, i.e., $(X^T, \zeta^T) \sim \boxed{Q_{X^T}} \otimes \boxed{P_{\zeta^T}}$

H_1 : X^T is LLM generated, i.e., $(X^T, \zeta^T) \sim \boxed{P_{X^T,\zeta^T}}$

Watermarking scheme

LLM Watermark Detection Errors

Watermark Detection $\Longrightarrow$ Hypothesis Testing: Human/unwatermarked LLM

$$\mathrm{H}_0 : X^T \text{ is human written, i.e., } (X^T, \zeta^T) \sim \boxed{Q_{X^T}} \otimes \boxed{P_{\zeta^T}}$$

$$\mathrm{H}_1 : X^T \text{ is LLM generated, i.e., } (X^T, \zeta^T) \sim \boxed{P_{X^T, \zeta^T}}$$

Watermarking scheme

Performance metric:

LLM Watermark Detection Errors

Watermark Detection $\Longrightarrow$ Hypothesis Testing: Human/unwatermarked LLM

$$\mathrm{H}_0 : X^T \text{ is human written, i.e., } (X^T, \zeta^T) \sim \boxed{Q_{X^T}} \otimes \boxed{P_{\zeta^T}}$$

$$\mathrm{H}_1 : X^T \text{ is LLM generated, i.e., } (X^T, \zeta^T) \sim \boxed{P_{X^T, \zeta^T}}$$

Watermarking scheme

Performance metric:

$$\boxed{\begin{matrix}\text{Detector} \\ \gamma\end{matrix}} \begin{cases} \mathrm{H}_0 : \text{Human} \\ \mathrm{H}_1 : \text{LLM} \end{cases}$$

LLM Watermark Detection Errors

Watermark Detection $\Longrightarrow$ Hypothesis Testing: Human/unwatermarked LLM

$$\mathrm{H}_0 : X^T \text{ is human written, i.e., } (X^T, \zeta^T) \sim \boxed{Q_{X^T}} \otimes \boxed{P_{\zeta^T}}$$

$$\mathrm{H}_1 : X^T \text{ is LLM generated, i.e., } (X^T, \zeta^T) \sim \boxed{P_{X^T, \zeta^T}}$$

Watermarking scheme

Performance metric:

Reality

Detector γ		H_0 : Human	H_1 : LLM
	H_0 : Human		
	H_1 : LLM		

LLM Watermark Detection Errors

Watermark Detection $\Longrightarrow$ Hypothesis Testing: Human/unwatermarked LLM

$$\mathrm{H}_0 : X^T \text{ is human written, i.e., } (X^T, \zeta^T) \sim \boxed{Q_{X^T}} \otimes \boxed{P_{\zeta^T}}$$

$$\mathrm{H}_1 : X^T \text{ is LLM generated, i.e., } (X^T, \zeta^T) \sim \boxed{P_{X^T,\zeta^T}}$$

Watermarking scheme

Performance metric:

Detector

γ

Reality

	H_0 : Human	H_1 : LLM
H_0 : Human	✅	
H_1 : LLM	False alarm $FA(\gamma, Q_{X^T}, P_{\zeta^T})$	

LLM Watermark Detection Errors

Watermark Detection $\Longrightarrow$ Hypothesis Testing:

Human/unwatermarked LLM

H_0 : X^T is human written, i.e., $(X^T, \zeta^T) \sim \boxed{Q_{X^T}} \otimes \boxed{P_{\zeta^T}}$

H_1 : X^T is LLM generated, i.e., $(X^T, \zeta^T) \sim \boxed{P_{X^T,\zeta^T}}$

Watermarking scheme

Performance metric:

Detector

γ

H_0 : Human

H_1 : LLM

Reality

	H_0 : Human	H_1 : LLM
H_0 : Human	✅	Miss detection $MD(\gamma, P_{X^T,\zeta^T})$
H_1 : LLM	False alarm $FA(\gamma, Q_{X^T}, P_{\zeta^T})$	✅

LLM Watermark Detection Errors

Watermark Detection $\Longrightarrow$ Hypothesis Testing:

Human/unwatermarked LLM

H_0 : X^T is human written, i.e., $(X^T, \zeta^T) \sim \boxed{Q_{X^T}} \otimes \boxed{P_{\zeta^T}}$

H_1 : X^T is LLM generated, i.e., $(X^T, \zeta^T) \sim \boxed{P_{X^T,\zeta^T}}$

Watermarking scheme

Performance metric:

Detector γ $\begin{cases} \mathrm{H}_0 : \text{Human} \\ \mathrm{H}_1 : \text{LLM} \end{cases}$

	Reality	
	H_0 : Human	H_1 : LLM
H_0 : Human	✅	Miss detection $\min\ MD(\gamma, P_{X^T,\zeta^T})$
H_1 : LLM	False alarm $FA(\gamma, Q_{X^T}, P_{\zeta^T}) \leq \alpha$	✅

LLM Watermarked Text Quality

Watermark Detection $\Longrightarrow$ Hypothesis Testing: Human/unwatermarked LLM

$$\mathrm{H}_0 : X^T \text{ is human written, i.e., } (X^T, \zeta^T) \sim \boxed{Q_{X^T}} \otimes \boxed{P_{\zeta^T}}$$

$$\mathrm{H}_1 : X^T \text{ is LLM generated, i.e., } (X^T, \zeta^T) \sim \boxed{P_{X^T, \zeta^T}}$$

Watermarking scheme

LLM Watermarked Text Quality

Watermark Detection ⟹ Hypothesis Testing:

Human/unwatermarked LLM

H_0 : X^T is human written, i.e., $(X^T, \zeta^T) \sim \boxed{Q_{X^T}} \otimes \boxed{P_{\zeta^T}}$

H_1 : X^T is LLM generated, i.e., $(X^T, \zeta^T) \sim \boxed{P_{X^T, \zeta^T}}$

Watermarking scheme

Other criteria for LLM watermarking?

LLM Watermarked Text Quality

Watermark Detection $\Longrightarrow$ Hypothesis Testing:

Human/unwatermarked LLM

$\mathrm{H}_0 : X^T$ is human written, i.e., $(X^T, \zeta^T) \sim \boxed{Q_{X^T}} \otimes \boxed{P_{\zeta^T}}$

$\mathrm{H}_1 : X^T$ is LLM generated, i.e., $(X^T, \zeta^T) \sim \boxed{P_{X^T, \zeta^T}}$

Watermarking scheme

Other criteria for LLM watermarking?

$\Longrightarrow$ **Text Quality!**

LLM Watermarked Text Quality

Watermark Detection $\Longrightarrow$ Hypothesis Testing:

Human/unwatermarked LLM

H_0 : X^T is human written, i.e., $(X^T, \zeta^T) \sim \boxed{Q_{X^T}} \otimes \boxed{P_{\zeta^T}}$

H_1 : X^T is LLM generated, i.e., $(X^T, \zeta^T) \sim \boxed{P_{X^T, \zeta^T}}$

Watermarking scheme

Other criteria for LLM watermarking?

$\Longrightarrow$ **Text Quality!**

$\Longrightarrow$ **Indistinguishable from unwatermarked**

LLM Watermarked Text Quality

Watermark Detection $\Longrightarrow$ Hypothesis Testing:

$\mathrm{H}_0 : X^T$ is human written, i.e., $(X^T, \zeta^T) \sim \boxed{Q_{X^T}} \otimes \boxed{P_{\zeta^T}}$

$\mathrm{H}_1 : X^T$ is LLM generated, i.e., $(X^T, \zeta^T) \sim \boxed{P_{X^T,\zeta^T}}$

Human/unwatermarked LLM

Watermarking scheme

watermarked text distribution

P_{X^T}

LLM Watermarked Text Quality

Watermark Detection $\Longrightarrow$ Hypothesis Testing:

Human/unwatermarked LLM

$\mathrm{H}_0 : X^T$ is human written, i.e., $(X^T, \zeta^T) \sim \boxed{Q_{X^T}} \otimes \boxed{P_{\zeta^T}}$

$\mathrm{H}_1 : X^T$ is LLM generated, i.e., $(X^T, \zeta^T) \sim \boxed{P_{X^T, \zeta^T}}$

Watermarking scheme

watermarked text distribution P_{X^T} **vs** original text distribution Q_{X^T}

LLM Watermarked Text Quality

Watermark Detection ⟹ Hypothesis Testing:

Human/unwatermarked LLM

$$\mathrm{H}_0 : X^T \text{ is human written, i.e., } (X^T, \zeta^T) \sim \boxed{Q_{X^T}} \otimes \boxed{P_{\zeta^T}}$$

$$\mathrm{H}_1 : X^T \text{ is LLM generated, i.e., } (X^T, \zeta^T) \sim \boxed{P_{X^T, \zeta^T}}$$

Watermarking scheme

watermarked text distribution P_{X^T} **vs** original text distribution Q_{X^T}

Good text quality

LLM Watermarked Text Quality

Watermark Detection $\Longrightarrow$ Hypothesis Testing:

Human/unwatermarked LLM

$$\mathrm{H}_0 : X^T \text{ is human written, i.e., } (X^T, \zeta^T) \sim \boxed{Q_{X^T}} \otimes \boxed{P_{\zeta^T}}$$

$$\mathrm{H}_1 : X^T \text{ is LLM generated, i.e., } (X^T, \zeta^T) \sim \boxed{P_{X^T, \zeta^T}}$$

Watermarking scheme

watermarked text distribution P_{X^T} **vs** original text distribution Q_{X^T}

Good text quality $\Longleftrightarrow$ $\mathrm{D}(P_{X^T}, Q_{X^T}) \leq \epsilon$

LLM Watermarked Text Quality

Watermark Detection $\Longrightarrow$ Hypothesis Testing:

Human/unwatermarked LLM

$$H_0 : X^T \text{ is human written, i.e., } (X^T, \zeta^T) \sim \boxed{Q_{X^T}} \otimes \boxed{P_{\zeta^T}}$$

$$H_1 : X^T \text{ is LLM generated, i.e., } (X^T, \zeta^T) \sim \boxed{P_{X^T, \zeta^T}}$$

Watermarking scheme

watermarked text distribution P_{X^T} **vs** original text distribution Q_{X^T}

Good text quality $\Longleftrightarrow$ $\mathrm{D}(P_{X^T}, Q_{X^T}) \leq \epsilon$

(Distortion Level)

LLM Watermarked Text Quality

Watermark Detection $\Longrightarrow$ Hypothesis Testing:

Human/unwatermarked LLM

$$\mathrm{H}_0 : X^T \text{ is human written, i.e., } (X^T, \zeta^T) \sim \boxed{Q_{X^T}} \otimes \boxed{P_{\zeta^T}}$$

$$\mathrm{H}_1 : X^T \text{ is LLM generated, i.e., } (X^T, \zeta^T) \sim \boxed{P_{X^T, \zeta^T}}$$

Watermarking scheme

watermarked text distribution P_{X^T} vs original text distribution Q_{X^T}

Good text quality $\Longleftrightarrow$ $\mathrm{D}(P_{X^T}, Q_{X^T}) \leq \epsilon$ (D can be any distortion metric)

(Distortion Level)

Trade-off in Designing LLM Watermarking

Watermark Detection $\Longrightarrow$ Hypothesis Testing:

Human/unwatermarked LLM

H_0 : X^T is human written, i.e., $(X^T, \zeta^T) \sim \boxed{Q_{X^T}} \otimes \boxed{P_{\zeta^T}}$

H_1 : X^T is LLM generated, i.e., $(X^T, \zeta^T) \sim \boxed{P_{X^T, \zeta^T}}$

Watermarking scheme

Trade-off in Designing LLM Watermarking

Watermark Detection $\Longrightarrow$ Hypothesis Testing:

Human/unwatermarked LLM

H_0 : X^T is human written, i.e., $(X^T, \zeta^T) \sim \boxed{Q_{X^T}} \otimes \boxed{P_{\zeta^T}}$

H_1 : X^T is LLM generated, i.e., $(X^T, \zeta^T) \sim \boxed{P_{X^T,\zeta^T}}$

Watermarking scheme

Trade-off:

Miss detection error, False alarm error, Distortion Level

Trade-off in Designing LLM Watermarking

Watermark Detection $\Longrightarrow$ Hypothesis Testing:

Human/unwatermarked LLM

$$\mathrm{H}_0 : X^T \text{ is human written, i.e., } (X^T, \zeta^T) \sim \boxed{Q_{X^T}} \otimes \boxed{P_{\zeta^T}}$$

$$\mathrm{H}_1 : X^T \text{ is LLM generated, i.e., } (X^T, \zeta^T) \sim \boxed{P_{X^T, \zeta^T}}$$

Watermarking scheme

Trade-off:

Miss detection error, False alarm error, Distortion Level

↑ ↓ ↓

Trade-off in Designing LLM Watermarking

Watermark Detection $\Longrightarrow$ Hypothesis Testing:

Human/unwatermarked LLM

$\mathrm{H}_0 : X^T$ is human written, i.e., $(X^T, \zeta^T) \sim \boxed{Q_{X^T}} \otimes \boxed{P_{\zeta^T}}$

$\mathrm{H}_1 : X^T$ is LLM generated, i.e., $(X^T, \zeta^T) \sim \boxed{P_{X^T, \zeta^T}}$

Watermarking scheme

Trade-off:

Miss detection error, False alarm error, Distortion Level

↓ ↑ ↑

Optimize LLM Watermark Generation and Detection

Watermark Detection $\Longrightarrow$ Hypothesis Testing: Human/unwatermarked LLM

$$\begin{aligned} \mathrm{H}_0 &: X^T \text{ is human written, i.e., } (X^T, \zeta^T) \sim \boxed{Q_{X^T}} \otimes \boxed{P_{\zeta^T}} \\ \mathrm{H}_1 &: X^T \text{ is LLM generated, i.e., } (X^T, \zeta^T) \sim \boxed{P_{X^T, \zeta^T}} \end{aligned}$$

Watermarking scheme

Optimize LLM Watermark Generation and Detection

Watermark Detection $\Longrightarrow$ Hypothesis Testing:

Human/unwatermarked LLM

H_0 : X^T is human written, i.e., $(X^T, \zeta^T) \sim \boxed{Q_{X^T}} \otimes \boxed{P_{\zeta^T}}$

H_1 : X^T is LLM generated, i.e., $(X^T, \zeta^T) \sim \boxed{P_{X^T,\zeta^T}}$

Watermarking scheme

Find the best watermarking scheme & detector:

Optimize LLM Watermark Generation and Detection

Watermark Detection $\Longrightarrow$ Hypothesis Testing: Human/unwatermarked LLM

$$\mathrm{H}_0 : X^T \text{ is human written, i.e., } (X^T, \zeta^T) \sim \boxed{Q_{X^T}} \otimes \boxed{P_{\zeta^T}}$$

$$\mathrm{H}_1 : X^T \text{ is LLM generated, i.e., } (X^T, \zeta^T) \sim \boxed{P_{X^T,\zeta^T}}$$

Watermarking scheme

<u>Find the best watermarking scheme & detector:</u>

Minimize miss detection →

$$\min_{\gamma,\, P_{X^T,\zeta^T}} \quad MD(\gamma,\, P_{X^T,\zeta^T})$$

Optimize LLM Watermark Generation and Detection

Watermark Detection $\Longrightarrow$ Hypothesis Testing:

Human/unwatermarked LLM

$\mathrm{H}_0 : X^T$ is human written, i.e., $(X^T, \zeta^T) \sim \boxed{Q_{X^T}} \otimes \boxed{P_{\zeta^T}}$

$\mathrm{H}_1 : X^T$ is LLM generated, i.e., $(X^T, \zeta^T) \sim \boxed{P_{X^T,\zeta^T}}$

Watermarking scheme

Find the best watermarking scheme & detector:

$$\min_{\gamma,\ P_{X^T,\zeta^T}} \quad MD(\gamma,\ P_{X^T,\zeta^T})$$

Humans are very creative, can write arbitrary texts

Optimize LLM Watermark Generation and Detection

Watermark Detection $\Longrightarrow$ Hypothesis Testing:

Human/unwatermarked LLM

$$\mathrm{H}_0 : X^T \text{ is human written, i.e., } (X^T, \zeta^T) \sim \boxed{Q_{X^T}} \otimes \boxed{P_{\zeta^T}}$$

$$\mathrm{H}_1 : X^T \text{ is LLM generated, i.e., } (X^T, \zeta^T) \sim \boxed{P_{X^T,\zeta^T}}$$

Watermarking scheme

Find the best watermarking scheme & detector:

$$\min_{\gamma,\, P_{X^T,\zeta^T}} \quad MD(\gamma,\, P_{X^T,\zeta^T})$$

$$\text{s.t.} \quad \sup_{Q_{X^T}} FA(\gamma, Q_{X^T}, P_{\zeta^T}) \le \alpha$$

Humans are very creative, can write arbitrary texts ➡

Optimize LLM Watermark Generation and Detection

Watermark Detection $\Longrightarrow$ Hypothesis Testing: Human/unwatermarked LLM

$$\mathrm{H}_0 : X^T \text{ is human written, i.e., } (X^T, \zeta^T) \sim \boxed{Q_{X^T}} \otimes \boxed{P_{\zeta^T}}$$

$$\mathrm{H}_1 : X^T \text{ is LLM generated, i.e., } (X^T, \zeta^T) \sim \boxed{P_{X^T,\zeta^T}}$$

Watermarking scheme

Find the best watermarking scheme & detector:

$$\min_{\gamma,\, P_{X^T,\zeta^T}} \quad MD(\gamma,\, P_{X^T,\zeta^T})$$

$$\text{s.t.} \quad \sup_{Q_{X^T}} FA(\gamma, Q_{X^T}, P_{\zeta^T}) \leq \alpha$$

Ensure text quality →

Optimize LLM Watermark Generation and Detection

Watermark Detection $\Longrightarrow$ Hypothesis Testing:

Human/unwatermarked LLM

H_0 : X^T is human written, i.e., $(X^T, \zeta^T) \sim \boxed{Q_{X^T}} \otimes \boxed{P_{\zeta^T}}$

H_1 : X^T is LLM generated, i.e., $(X^T, \zeta^T) \sim \boxed{P_{X^T,\zeta^T}}$

Watermarking scheme

Find the best watermarking scheme & detector:

$$\min_{\gamma,\, P_{X^T,\zeta^T}} \quad MD(\gamma,\, P_{X^T,\zeta^T})$$

$$\text{s.t.} \quad \sup_{Q_{X^T}} FA(\gamma, Q_{X^T}, P_{\zeta^T}) \le \alpha$$

Ensure text quality ➡ $\mathrm{D}(P_{X^T}, Q_{X^T}) \le \epsilon$

Optimize LLM Watermark Generation and Detection

Watermark Detection $\Longrightarrow$ Hypothesis Testing:

Human/unwatermarked LLM

$$\mathrm{H}_0 : X^T \text{ is human written, i.e., } (X^T, \zeta^T) \sim \boxed{Q_{X^T}} \otimes \boxed{P_{\zeta^T}}$$

$$\mathrm{H}_1 : X^T \text{ is LLM generated, i.e., } (X^T, \zeta^T) \sim \boxed{P_{X^T,\zeta^T}}$$

Watermarking scheme

Find the best watermarking scheme & detector:

$$\min_{\gamma,\, P_{X^T,\zeta^T}} \quad MD(\gamma,\ P_{X^T,\zeta^T})$$

$$\text{s.t.} \quad \sup_{Q_{X^T}} FA(\gamma, Q_{X^T}, P_{\zeta^T}) \leq \alpha$$

$$\mathrm{D}(P_{X^T}, Q_{X^T}) \leq \epsilon$$

Fundamental Limit for Miss Detection Error

Optimization problem:

$$\min_{\gamma,\, P_{X^T,\zeta^T}} \quad MD(\gamma, P_{X^T,\zeta^T})$$

$$\text{s.t.} \quad \sup_{Q_{X^T}} FA(\gamma, Q_{X^T}, P_{\zeta^T}) \leq \alpha$$

$$\mathrm{D}(P_{X^T}, Q_{X^T}) \leq \epsilon$$

Fundamental Limit for Miss Detection Error

Watermarked text distribution: $P^*_{X^T} = \arg \min_{P_{X^T}:\mathrm{D}(P_{X^T}, Q_{X^T}) \le \epsilon} \sum_{x^T} (P_{X^T}(x^T) - \alpha)_+$

Optimization problem:

$$\min_{\gamma,\, P_{X^T,\zeta^T}} \quad MD(\gamma,\, P_{X^T,\zeta^T})$$

s.t. $\sup_{Q_{X^T}} FA(\gamma, Q_{X^T}, P_{\zeta^T}) \le \alpha$

$\mathrm{D}(P_{X^T}, Q_{X^T}) \le \epsilon$

✦ **Minimum miss detection error:**

$$MD^*(Q_{X^T}, \alpha, \epsilon) = \sum_{x^T} \left(P^*_{X^T}(x^T) - \alpha\right)_+$$

Fundamental Limit for Miss Detection Error

Watermarked text distribution: $P^*_{X^T} = \arg \min_{P_{X^T}:\mathsf{D}(P_{X^T}, Q_{X^T}) \leq \epsilon} \sum_{x^T} (P_{X^T}(x^T) - \alpha)_+$

Optimization problem:

$$\min_{\gamma,\, P_{X^T,\zeta^T}} \quad MD(\gamma, P_{X^T,\zeta^T})$$

$$\text{s.t.} \quad \sup_{Q_{X^T}} FA(\gamma, Q_{X^T}, P_{\zeta^T}) \leq \alpha$$

$$\mathsf{D}(P_{X^T}, Q_{X^T}) \leq \epsilon$$

✦ **Minimum miss detection error:**

$$MD^*(Q_{X^T}, \alpha, \epsilon) = \sum_{x^T} \left(P^*_{X^T}(x^T) - \alpha\right)_+$$

Best achievable for any watermarking methods

Fundamental Limit for Miss Detection Error

Watermarked text distribution: $P^*_{X^T} = \arg \min_{P_{X^T}:\mathrm{D}(P_{X^T}, Q_{X^T}) \leq \epsilon} \sum_{x^T} (P_{X^T}(x^T) - \alpha)_+$

<u>Optimization problem:</u>

$$\min_{\gamma,\, P_{X^T,\zeta^T}} \quad MD(\gamma,\, P_{X^T,\zeta^T})$$

$$\text{s.t.} \quad \sup_{Q_{X^T}} FA(\gamma, Q_{X^T}, P_{\zeta^T}) \leq \alpha$$

$$\mathrm{D}(P_{X^T}, Q_{X^T}) \leq \epsilon$$

✦ **Minimum miss detection error:**

$$MD^*(Q_{X^T}, \alpha, \epsilon) = \sum_{x^T} \left(P^*_{X^T}(x^T) - \alpha\right)_+$$

Fundamental Limit for Miss Detection Error

Watermarked text distribution: $P^*_{X^T} = \arg \min_{P_{X^T}:\mathsf{D}(P_{X^T}, Q_{X^T}) \leq \epsilon} \sum_{x^T} (P_{X^T}(x^T) - \alpha)_+$

Optimization problem:

$$\min_{\gamma, P_{X^T,\zeta^T}} \quad MD(\gamma, P_{X^T,\zeta^T})$$

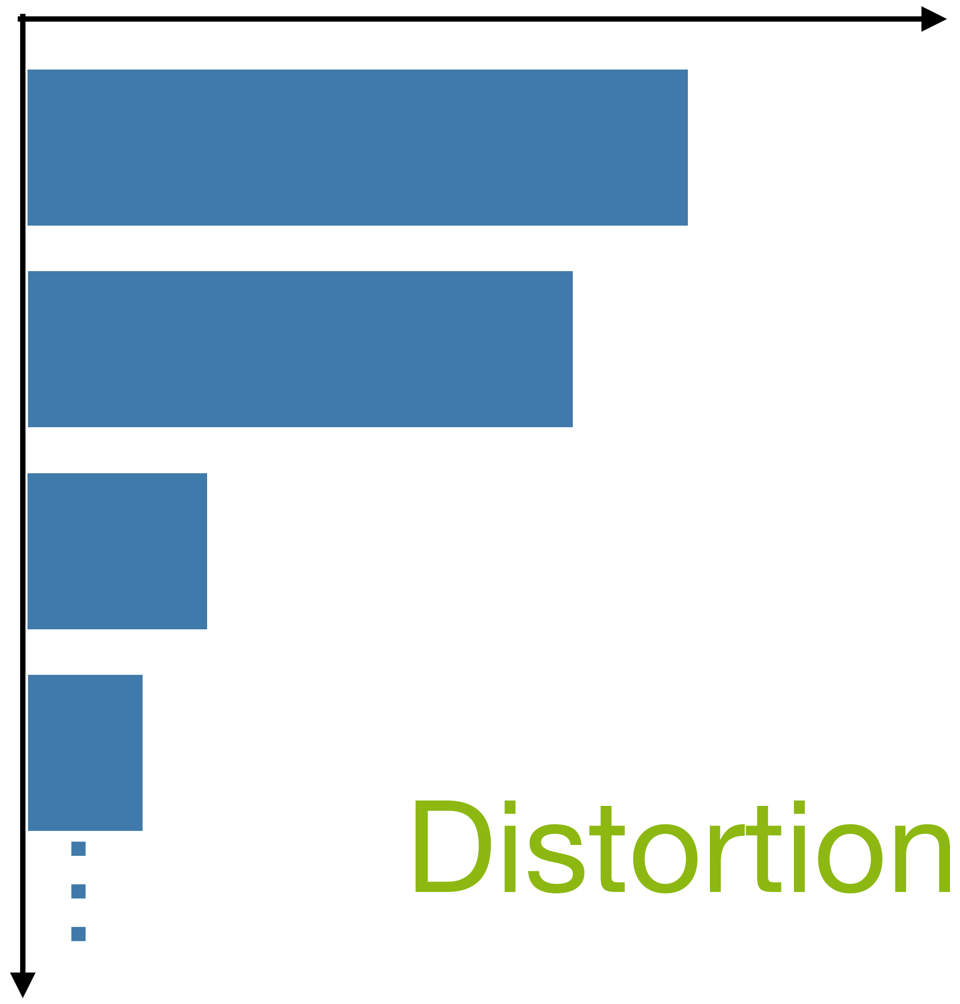
$$\text{s.t.} \quad \sup_{Q_{X^T}} FA(\gamma, Q_{X^T}, P_{\zeta^T}) \leq \alpha$$

$$\mathsf{D}(P_{X^T}, Q_{X^T}) \leq \epsilon$$

✦ **Minimum miss detection error:**

$$MD^*(Q_{X^T}, \alpha, \epsilon) = \sum_{x^T} \left(P^*_{X^T}(x^T) - \alpha\right)_+$$

$P^*_{X^T} = Q_{X^T}$



Distortion-free

Fundamental Limit for Miss Detection Error

Watermarked text distribution: $P^*_{X^T} = \arg \min_{P_{X^T}:\mathsf{D}(P_{X^T}, Q_{X^T}) \le \epsilon} \sum_{x^T} (P_{X^T}(x^T) - \alpha)_+$

Optimization problem:

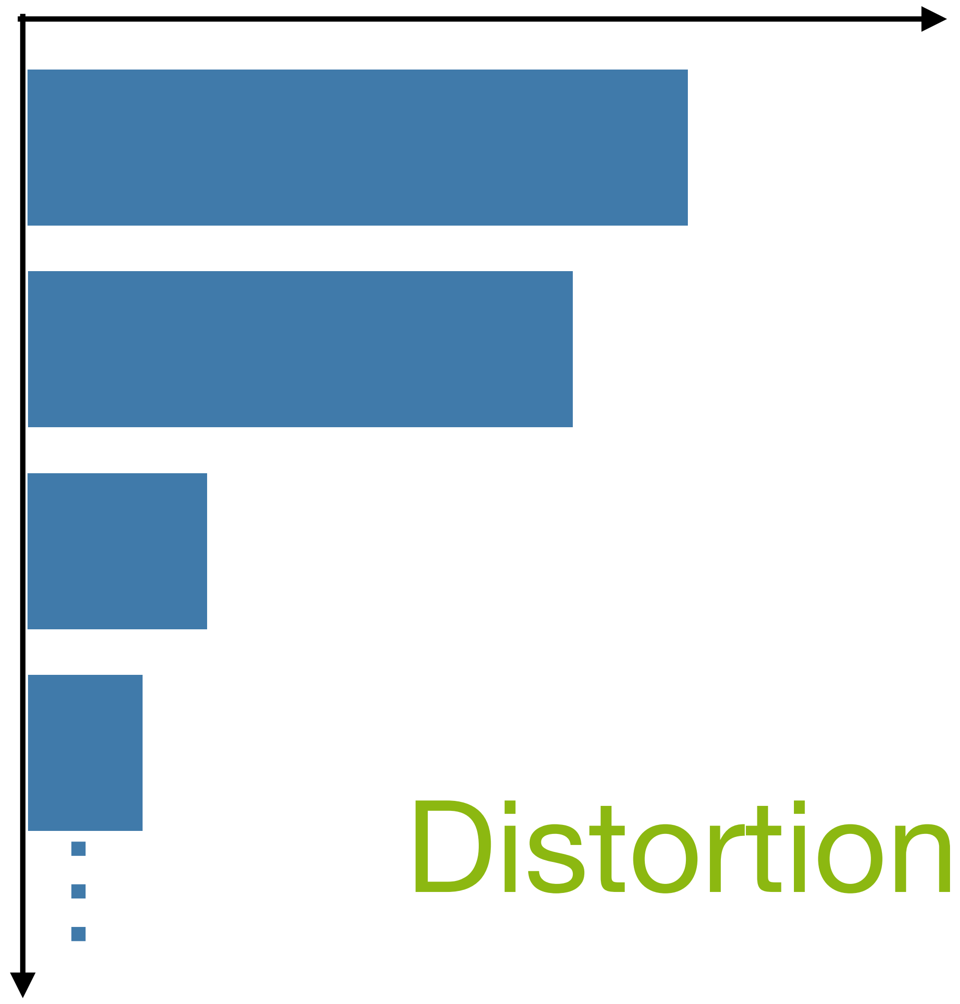
$$\begin{aligned} \min_{\gamma,\, P_{X^T,\zeta^T}} \quad & MD(\gamma, P_{X^T,\zeta^T}) \\ \text{s.t.} \quad & \sup_{Q_{X^T}} FA(\gamma, Q_{X^T}, P_{\zeta^T}) \le \alpha \\ & \mathsf{D}(P_{X^T}, Q_{X^T}) \le \epsilon \end{aligned}$$

D_{TV}

✦ **Minimum miss detection error:**

$$MD^*(Q_{X^T}, \alpha, \epsilon) = \sum_{x^T} \left(P^*_{X^T}(x^T) - \alpha \right)_+$$

$P^*_{X^T} = Q_{X^T}$



Distortion-free

Fundamental Limit for Miss Detection Error

Watermarked text distribution: $P^*_{X^T} = \arg \min_{P_{X^T}:\mathsf{D}(P_{X^T}, Q_{X^T}) \leq \epsilon} \sum_{x^T} (P_{X^T}(x^T) - \alpha)_+$

Optimization problem:

$$\min_{\gamma,\, P_{X^T,\zeta^T}} \quad MD(\gamma,\, P_{X^T,\zeta^T})$$

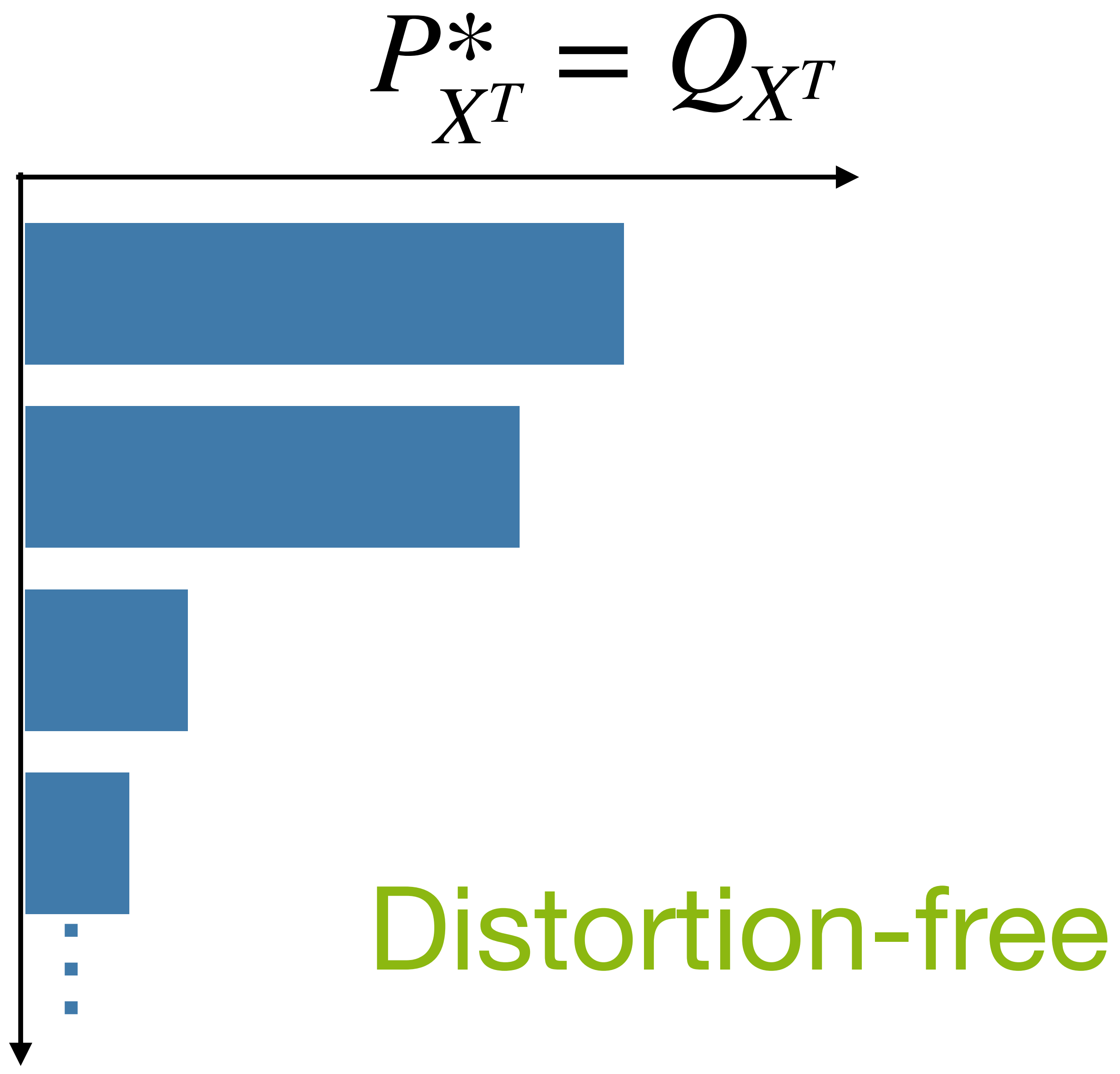
$$\text{s.t.} \quad \sup_{Q_{X^T}} FA(\gamma, Q_{X^T}, P_{\zeta^T}) \leq \alpha$$

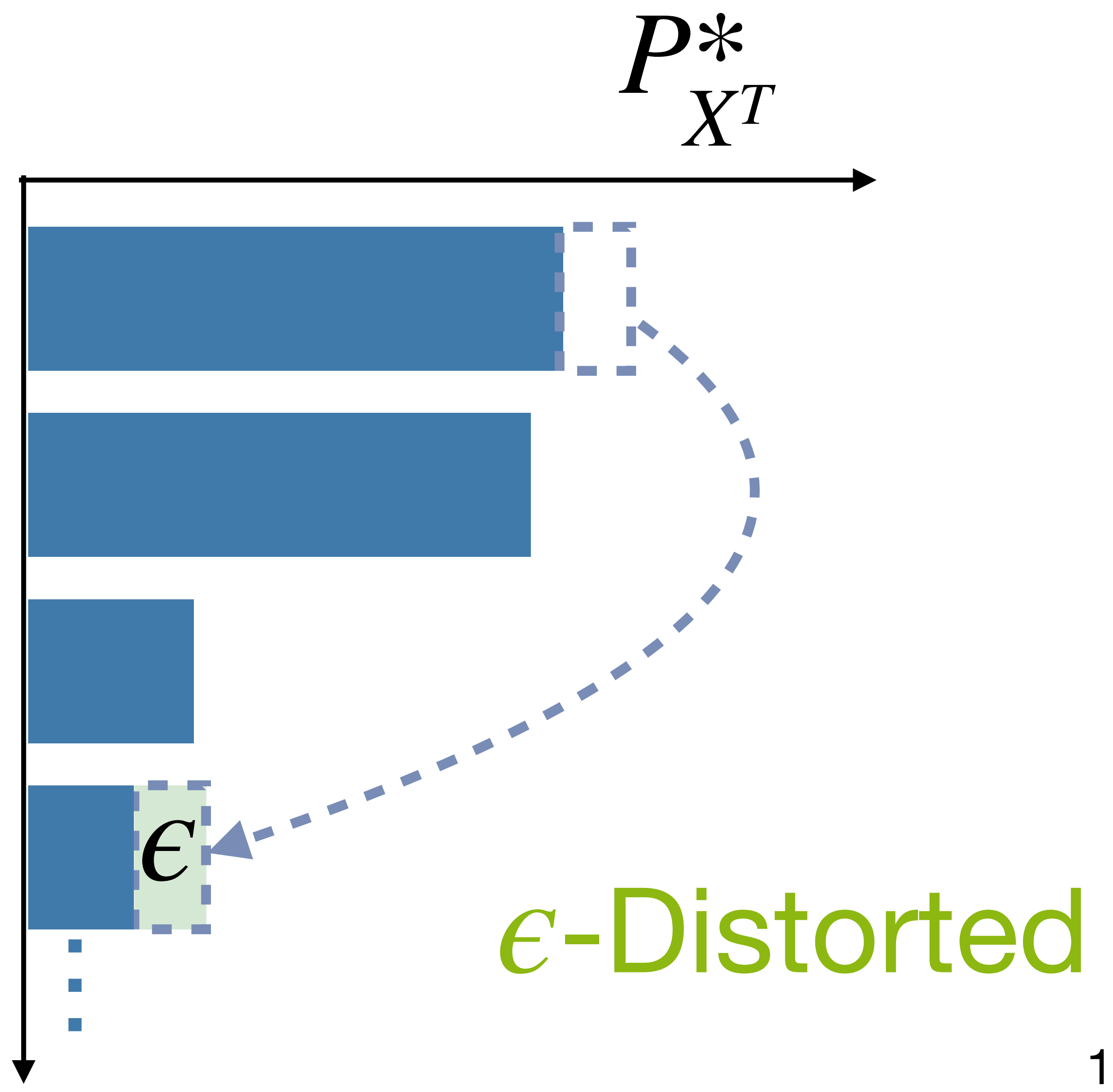
$$\mathsf{D}(P_{X^T}, Q_{X^T}) \leq \epsilon$$

D_{TV}

✦ **Minimum miss detection error:**

$$MD^*(Q_{X^T}, \alpha, \epsilon) = \sum_{x^T} \left(P^*_{X^T}(x^T) - \alpha\right)_+$$





Fundamental Limit for Miss Detection Error

Watermarked text distribution: $P^*_{X^T} = \arg \min_{P_{X^T}:\mathsf{D}(P_{X^T}, Q_{X^T}) \le \epsilon} \sum_{x^T} (P_{X^T}(x^T) - \alpha)_+$

Optimization problem:

$$\min_{\gamma,\, P_{X^T,\zeta^T}} \quad MD(\gamma,\, P_{X^T,\zeta^T})$$

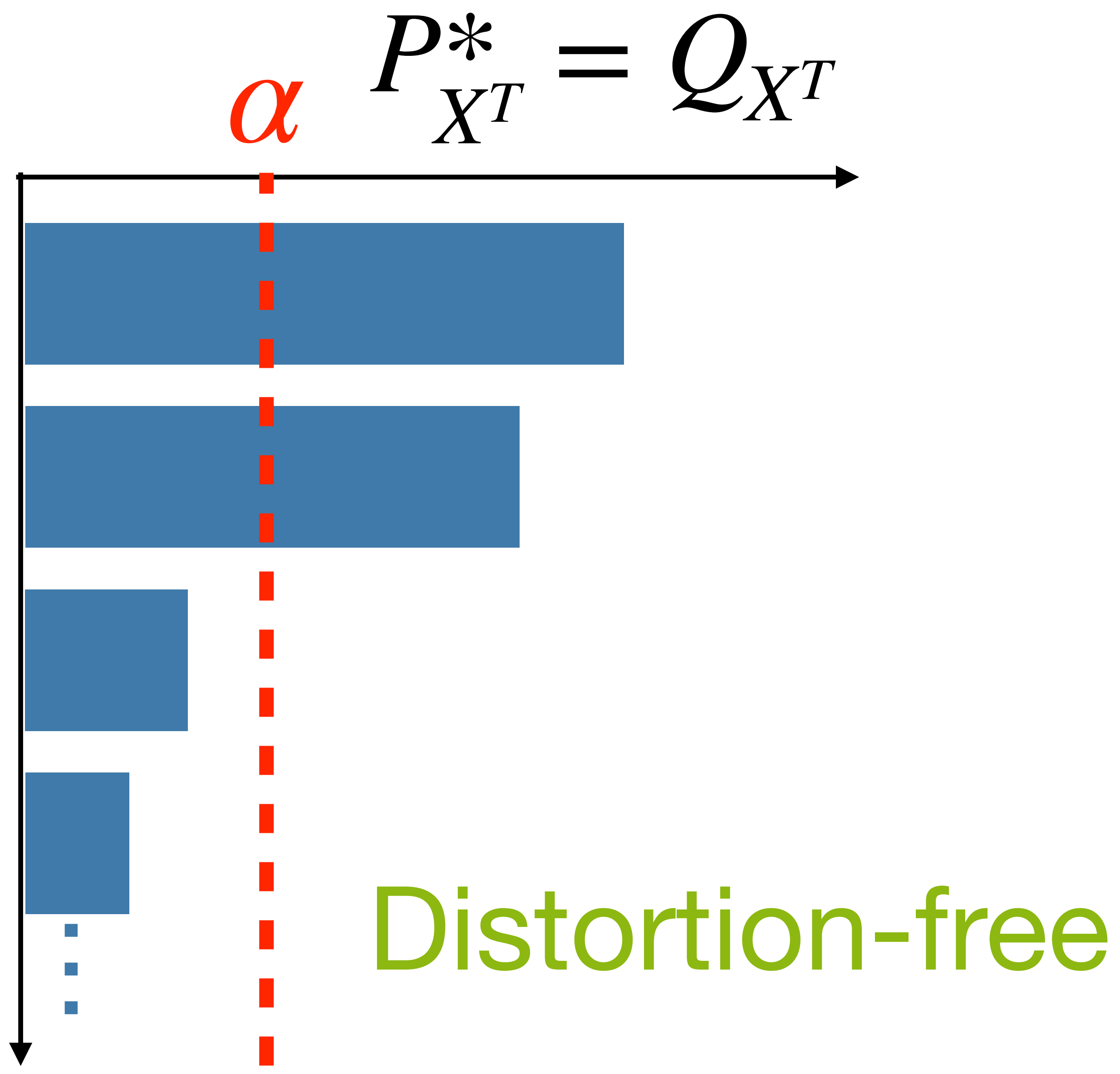
$$\text{s.t.} \quad \sup_{Q_{X^T}} FA(\gamma, Q_{X^T}, P_{\zeta^T}) \le \alpha$$

$$\mathsf{D}(P_{X^T}, Q_{X^T}) \le \epsilon$$

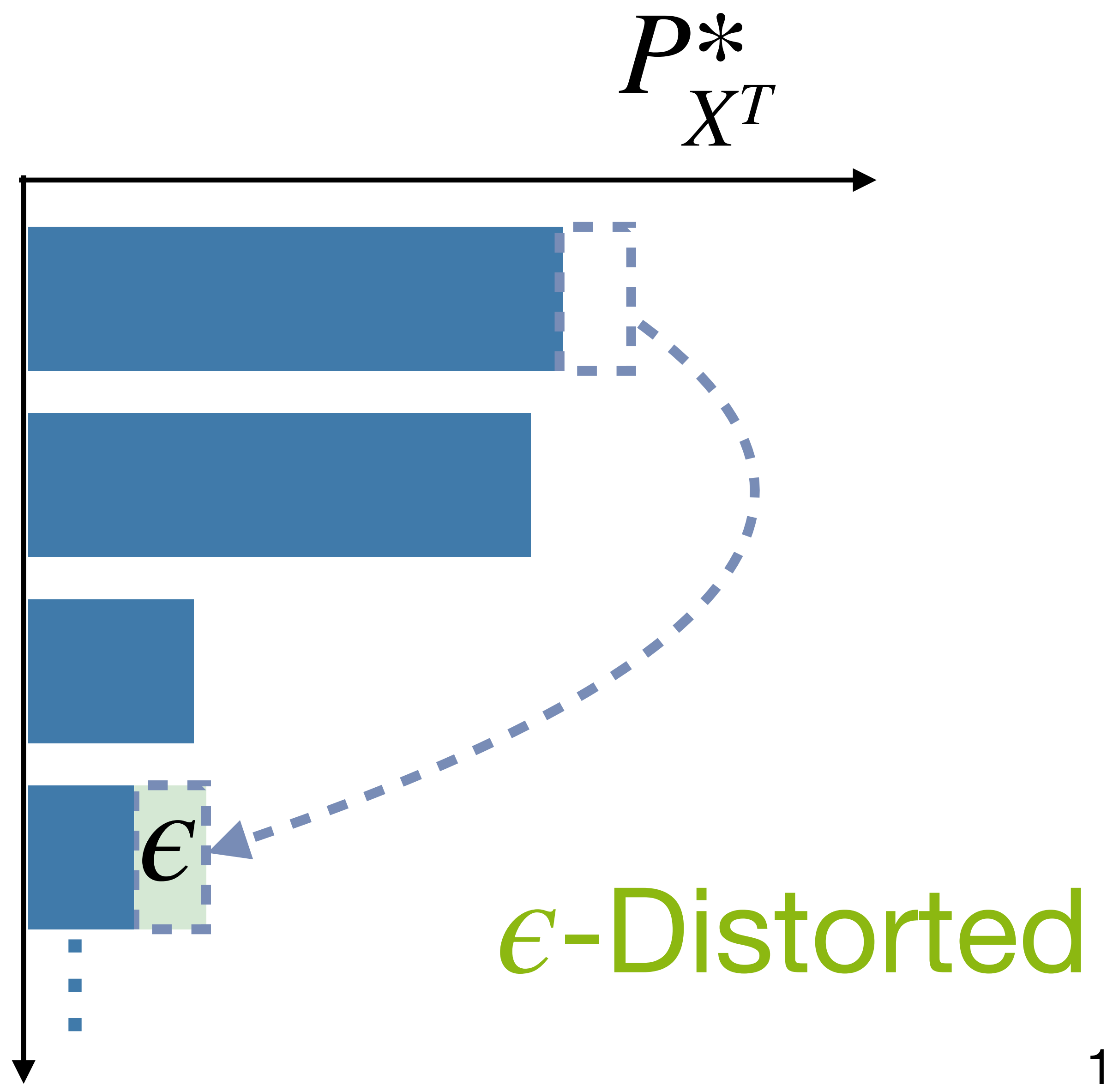
D_{TV}

✦ **Minimum miss detection error:**

$$MD^*(Q_{X^T}, \alpha, \epsilon) = \sum_{x^T} \left(P^*_{X^T}(x^T) - \alpha \right)_+$$



Distortion-free



ϵ-Distorted

Fundamental Limit for Miss Detection Error

Watermarked text distribution: $P^*_{X^T} = \arg\min_{P_{X^T}:\mathsf{D}(P_{X^T}, Q_{X^T}) \leq \epsilon} \sum_{x^T} (P_{X^T}(x^T) - \alpha)_+$

Optimization problem:

$$\min_{\gamma,\, P_{X^T,\zeta^T}} \quad MD(\gamma, P_{X^T,\zeta^T})$$

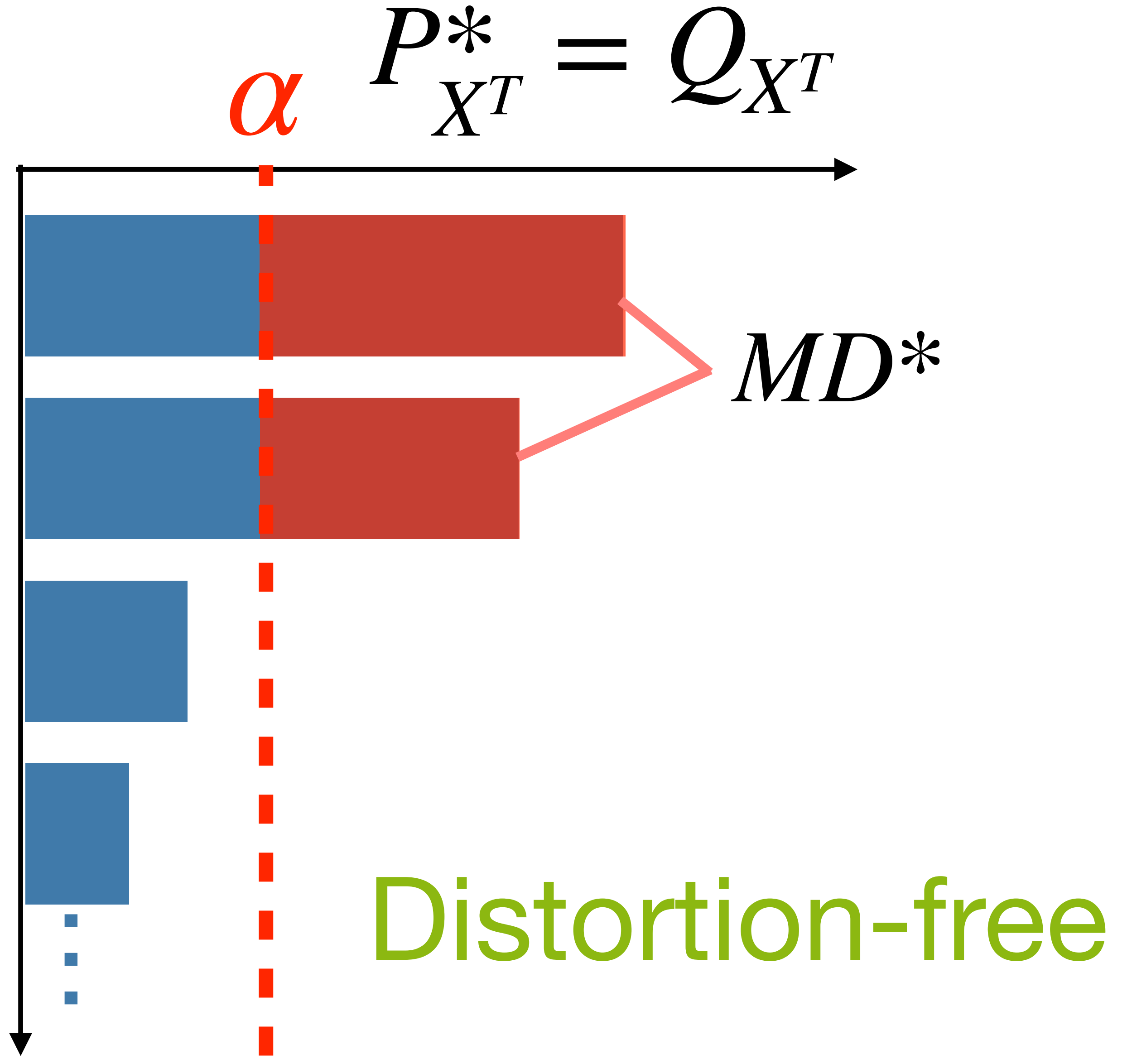
$$\text{s.t.} \quad \sup_{Q_{X^T}} FA(\gamma, Q_{X^T}, P_{\zeta^T}) \leq \alpha$$

$$\mathsf{D}(P_{X^T}, Q_{X^T}) \leq \epsilon$$

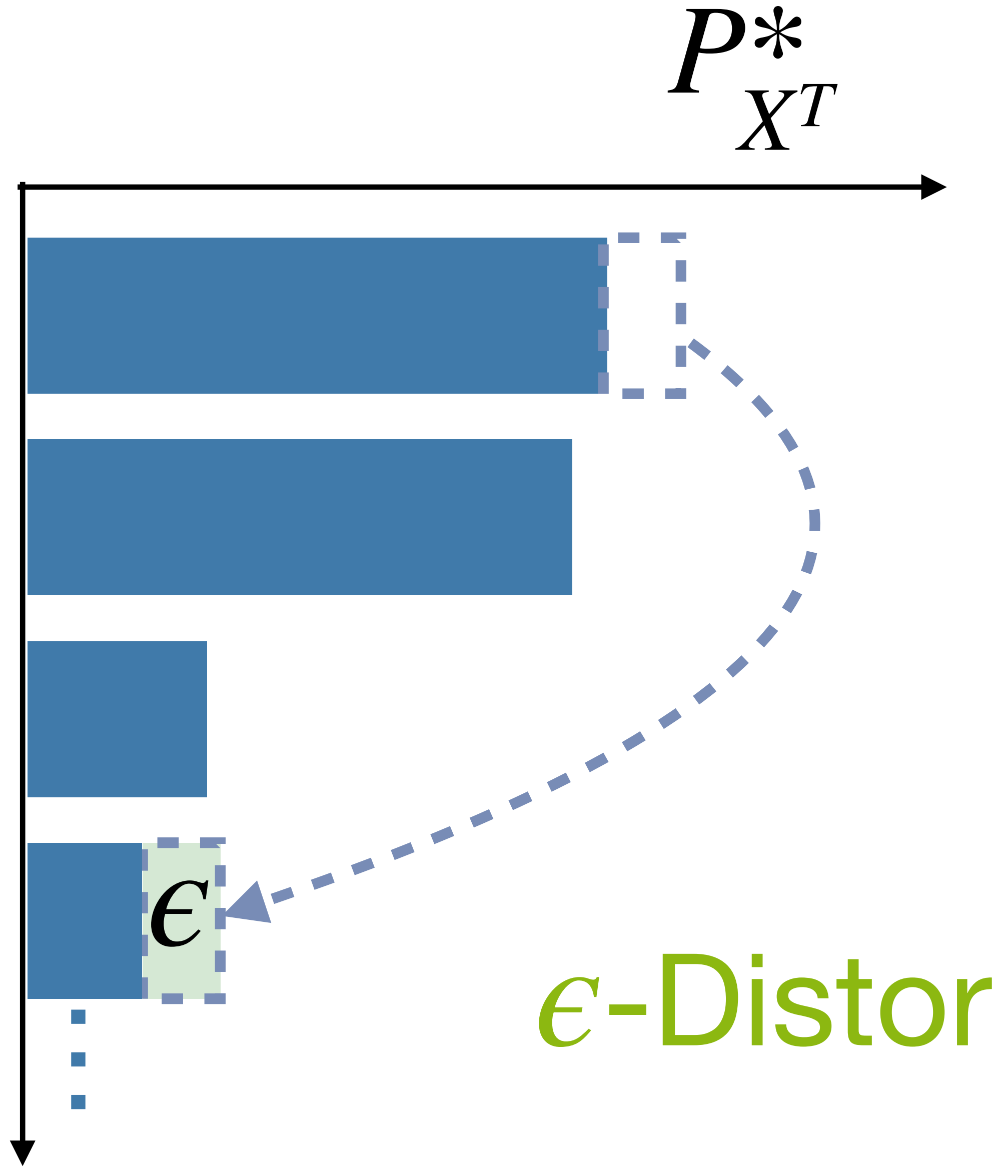
D_{TV}

✦ Minimum miss detection error:

$$MD^*(Q_{X^T}, \alpha, \epsilon) = \sum_{x^T} \left(P^*_{X^T}(x^T) - \alpha\right)_+$$



Distortion-free



ϵ-Distorted

Fundamental Limit for Miss Detection Error

Watermarked text distribution: $P^*_{X^T} = \arg \min_{P_{X^T}: \mathsf{D}(P_{X^T}, Q_{X^T}) \le \epsilon} \sum_{x^T} (P_{X^T}(x^T) - \alpha)_+$

Optimization problem:

$$\min_{\gamma,\, P_{X^T,\zeta^T}} \quad MD(\gamma, P_{X^T,\zeta^T})$$

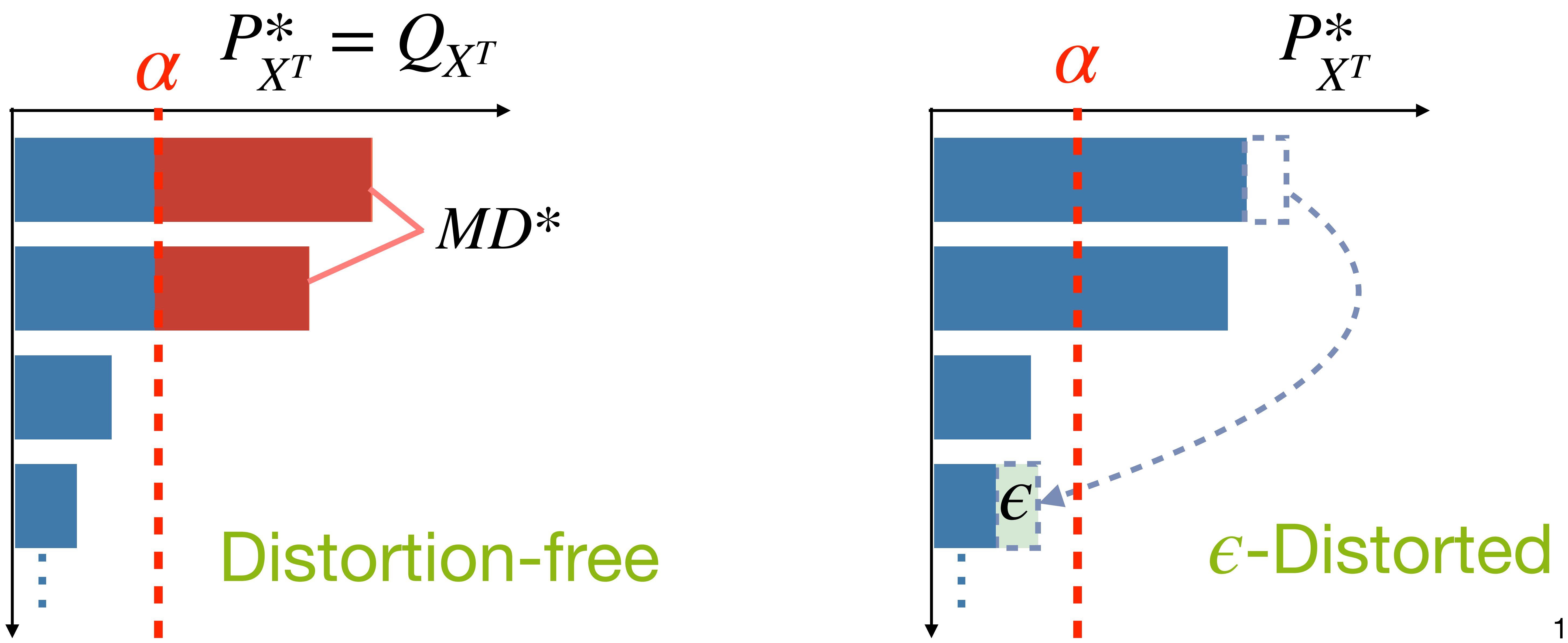
s.t. $\sup_{Q_{X^T}} FA(\gamma, Q_{X^T}, P_{\zeta^T}) \le \alpha$

$\mathsf{D}(P_{X^T}, Q_{X^T}) \le \epsilon$

D_{TV}

✦ **Minimum miss detection error:**

$$MD^*(Q_{X^T}, \alpha, \epsilon) = \sum_{x^T} \left(P^*_{X^T}(x^T) - \alpha\right)_+$$



Fundamental Limit for Miss Detection Error

Watermarked text distribution: $P^*_{X^T} = \arg \min_{P_{X^T}: \mathrm{D}(P_{X^T}, Q_{X^T}) \le \epsilon} \sum_{x^T} (P_{X^T}(x^T) - \alpha)_+$

Optimization problem:

$$\min_{\gamma,\, P_{X^T,\zeta^T}} \quad MD(\gamma,\, P_{X^T,\zeta^T})$$

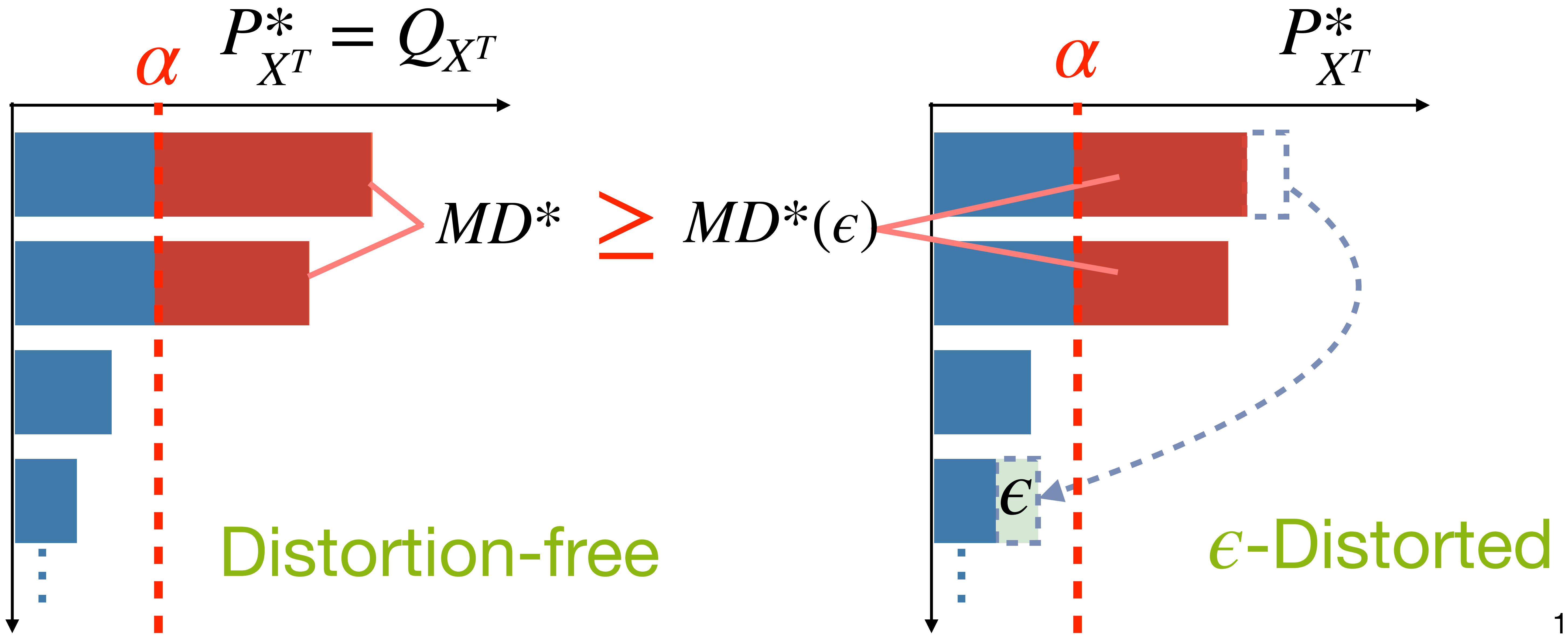
$$\text{s.t.} \quad \sup_{Q_{X^T}} FA(\gamma, Q_{X^T}, P_{\zeta^T}) \le \alpha$$

$$\mathrm{D}(P_{X^T}, Q_{X^T}) \le \epsilon$$

D_{TV}

✦ Minimum miss detection error:

$$MD^*(Q_{X^T}, \alpha, \epsilon) = \sum_{x^T} \left(P^*_{X^T}(x^T) - \alpha\right)_+$$



Jointly Optimal Detector and Watermarking Scheme

Optimization problem:

$$\min_{\gamma,\, P_{X^T,\zeta^T}} \quad MD(\gamma,\, P_{X^T,\zeta^T})$$

$$\text{s.t.} \quad \sup_{Q_{X^T}} FA(\gamma, Q_{X^T}, P_{\zeta^T}) \leq \alpha$$

$$\mathsf{D}(P_{X^T}, Q_{X^T}) \leq \epsilon$$

Jointly Optimal Detector and Watermarking Scheme

Optimization problem:

$$\min_{\gamma,\, P_{X^T,\zeta^T}} \quad MD(\gamma,\, P_{X^T,\zeta^T})$$

$$\text{s.t.} \quad \sup_{Q_{X^T}} FA(\gamma, Q_{X^T}, P_{\zeta^T}) \leq \alpha$$

$$\mathrm{D}(P_{X^T}, Q_{X^T}) \leq \epsilon$$

✦ **Jointly optimal detector γ^* and watermarking scheme $P^*_{X^T,\zeta^T}$:**

Jointly Optimal Detector and Watermarking Scheme

Optimization problem:

$$\min_{\gamma,\, P_{X^T,\zeta^T}} \quad MD(\gamma,\, P_{X^T,\zeta^T})$$

$$\text{s.t.} \quad \sup_{Q_{X^T}} FA(\gamma, Q_{X^T}, P_{\zeta^T}) \leq \alpha$$

$$\mathrm{D}(P_{X^T}, Q_{X^T}) \leq \epsilon$$

✦ **Jointly optimal detector γ^* and watermarking scheme $P^*_{X^T,\zeta^T}$:**

$$\gamma^* = \mathbf{1}\{X^T = g(\zeta^T)\}$$

for some surjective $g : \mathcal{Z}^T \to \mathcal{S} \supset \mathcal{V}^T$

Jointly Optimal Detector and Watermarking Scheme

Optimization problem:

$$\min_{\gamma,\, P_{X^T,\zeta^T}} \quad MD(\gamma,\, P_{X^T,\zeta^T})$$

$$\text{s.t.} \quad \sup_{Q_{X^T}} FA(\gamma, Q_{X^T}, P_{\zeta^T}) \le \alpha$$

$$\mathrm{D}(P_{X^T}, Q_{X^T}) \le \epsilon$$

✦ **Jointly optimal detector γ^* and watermarking scheme $P^*_{X^T,\zeta^T}$:**

$$\gamma^* = \mathbf{1}\{X^T = g(\zeta^T)\}$$

for some surjective $g : \mathscr{Z}^T \to \mathscr{S} \supset \mathscr{V}^T$

$P^*_{X^T,\zeta^T}$:

Jointly Optimal Detector and Watermarking Scheme

Optimization problem:

$$\min_{\gamma,\, P_{X^T,\zeta^T}} \quad MD(\gamma,\, P_{X^T,\zeta^T})$$

$$\text{s.t.} \quad \sup_{Q_{X^T}} FA(\gamma, Q_{X^T}, P_{\zeta^T}) \leq \alpha \quad (\Delta_1)$$

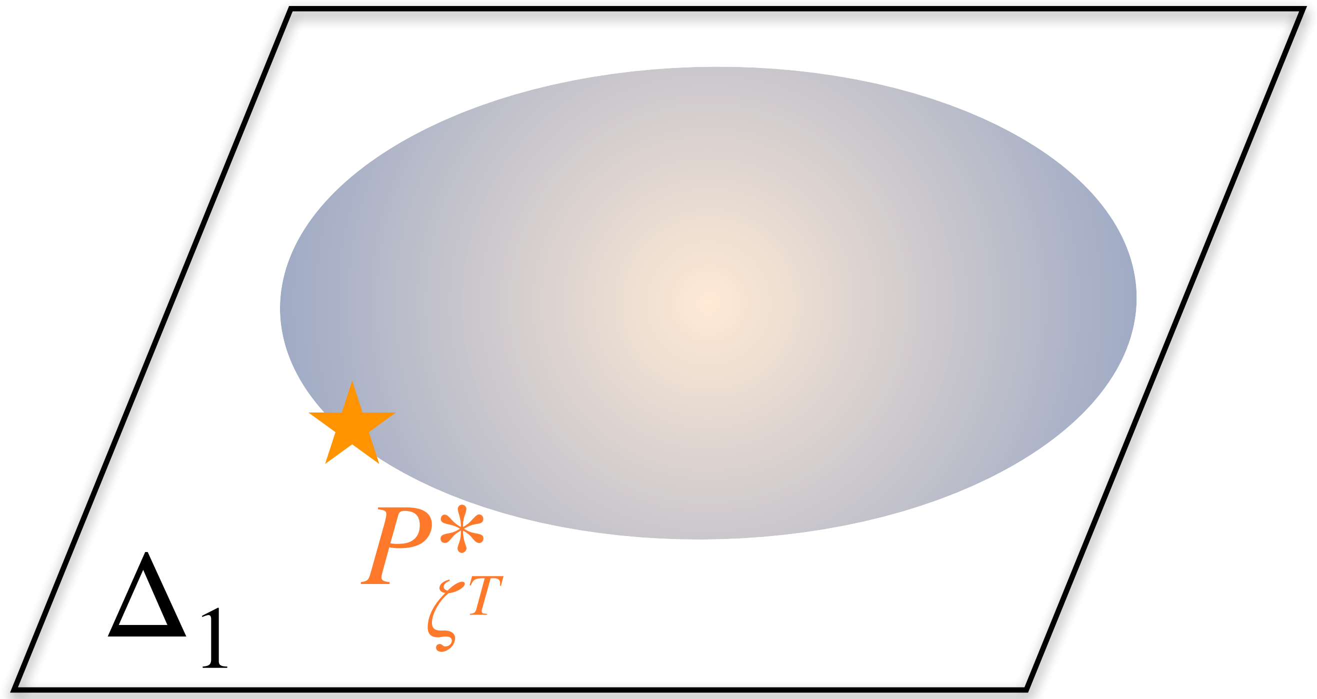
$$\mathrm{D}(P_{X^T}, Q_{X^T}) \leq \epsilon$$

✦ **Jointly optimal detector γ^* and watermarking scheme $P^*_{X^T,\zeta^T}$:**

$$\gamma^* = \mathbf{1}\{X^T = g(\zeta^T)\}$$

for some surjective $g : \mathscr{Z}^T \to \mathscr{S} \supset \mathscr{V}^T$

$P^*_{X^T,\zeta^T}$:



Jointly Optimal Detector and Watermarking Scheme

Optimization problem:

$$\min_{\gamma,\, P_{X^T,\zeta^T}} \quad MD(\gamma, P_{X^T,\zeta^T})$$

$$\text{s.t.} \quad \sup_{Q_{X^T}} FA(\gamma, Q_{X^T}, P_{\zeta^T}) \le \alpha \qquad (\Delta_1)$$

$$\mathrm{D}(P_{X^T}, Q_{X^T}) \le \epsilon \qquad (\Delta_2)$$

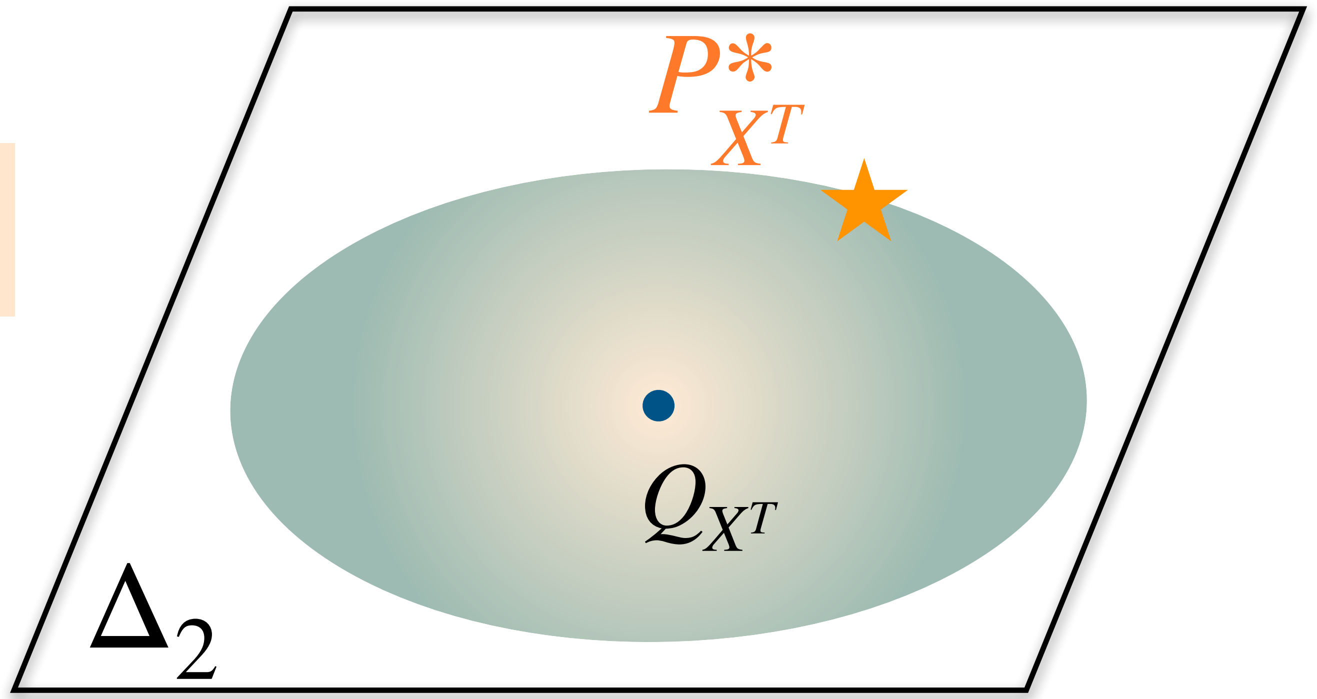
$$P^*_{X^T} = \arg \min_{P_{X^T}:\mathrm{D}(P_{X^T}, Q_{X^T}) \le \epsilon} \sum_{x^T} (P_{X^T}(x^T) - \alpha)_+$$

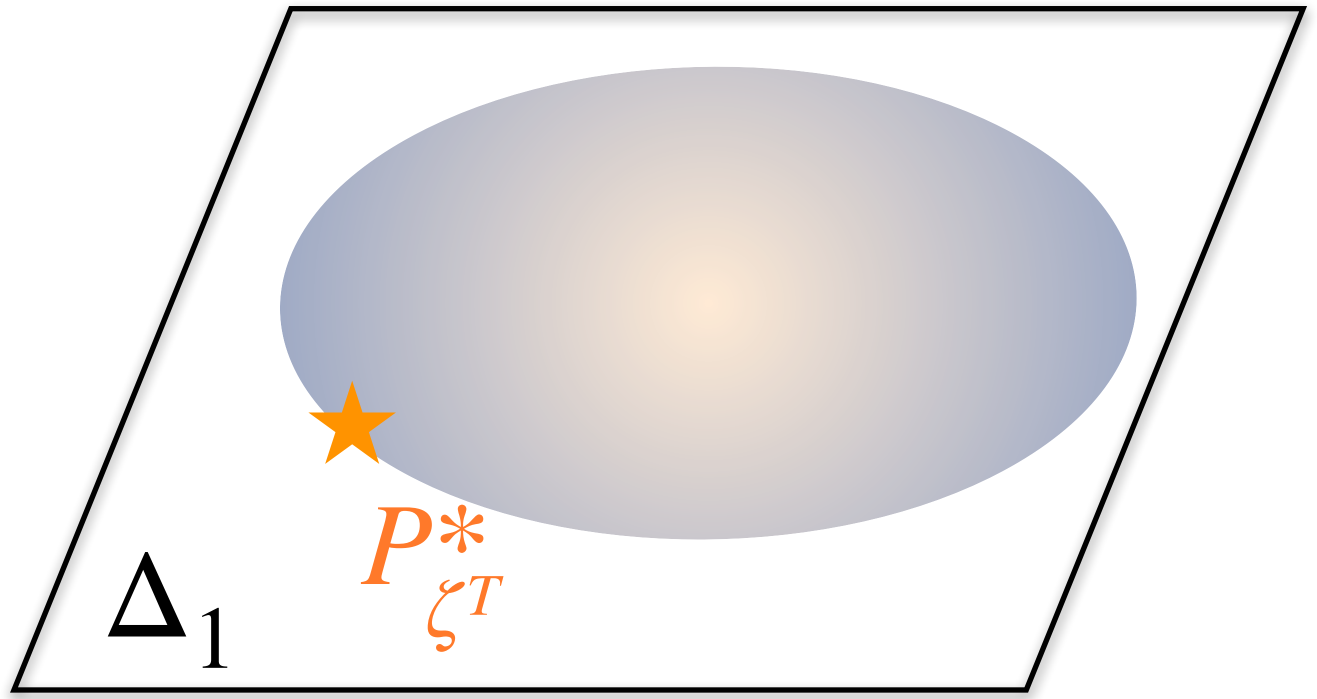
✦ **Jointly optimal detector γ^* and watermarking scheme $P^*_{X^T,\zeta^T}$:**

$$\gamma^* = \mathbf{1}\{X^T = g(\zeta^T)\}$$

for some surjective $g : \mathscr{Z}^T \to \mathscr{S} \supset \mathscr{V}^T$

$P^*_{X^T,\zeta^T}$:





Jointly Optimal Detector and Watermarking Scheme

Optimization problem:

$$\min_{\gamma,\, P_{X^T,\zeta^T}} \quad MD(\gamma, P_{X^T,\zeta} = \mathbb{E}_{P_{X^T,\zeta^T}}[1-\gamma(X^T,\zeta^T)]$$

$$\text{s.t.} \quad \sup_{Q_{X^T}} FA(\gamma, Q_{X^T}, P_{\zeta^T}) \le \alpha \qquad (\Delta_1)$$

$$\mathsf{D}(P_{X^T}, Q_{X^T}) \le \epsilon \qquad (\Delta_2)$$

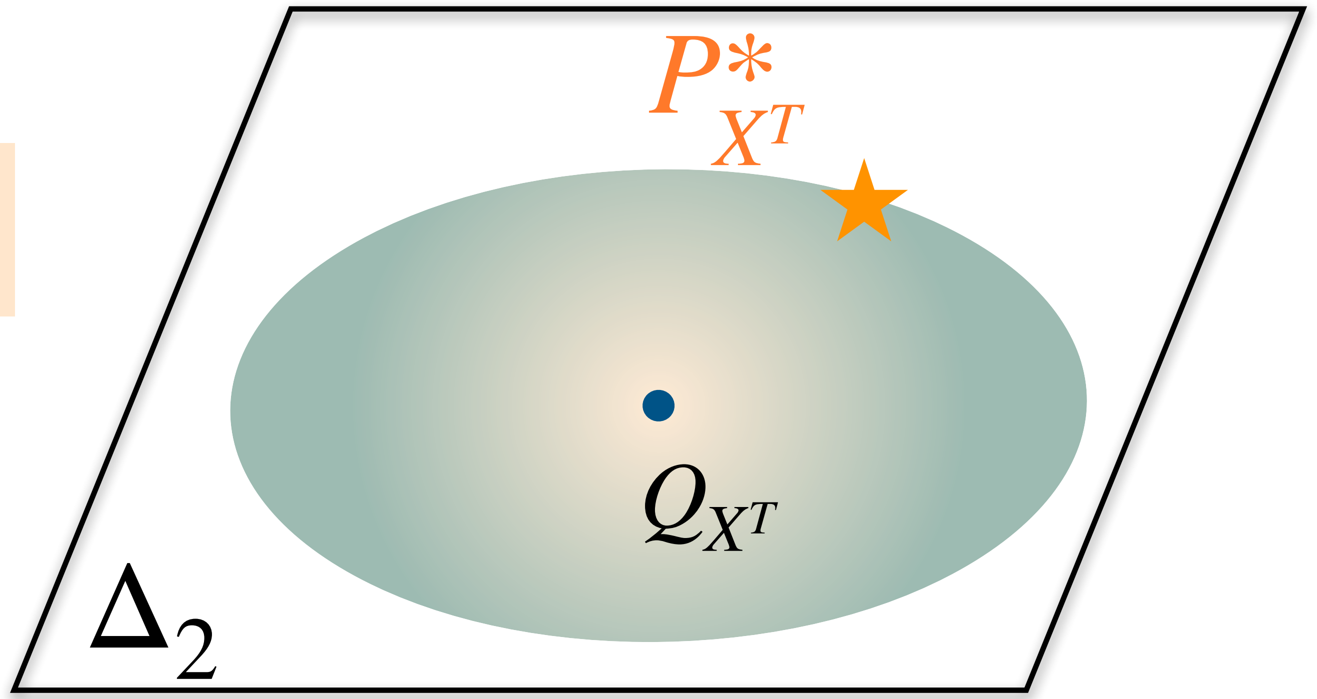
$$P^*_{X^T} = \arg\min_{P_{X^T}:\mathsf{D}(P_{X^T},Q_{X^T})\le\epsilon} \sum_{x^T} (P_{X^T}(x^T)-\alpha)_+$$

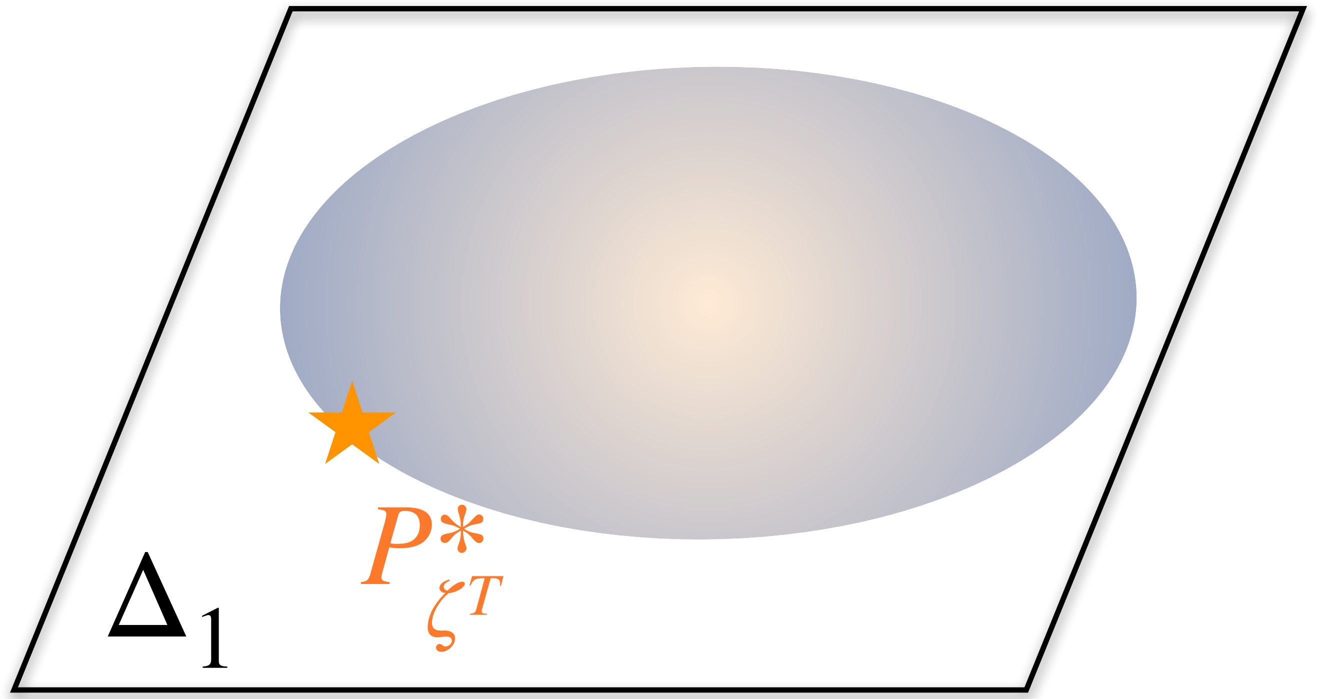
✦ **Jointly optimal detector γ^* and watermarking scheme $P^*_{X^T,\zeta^T}$:**

$$\gamma^* = \mathbf{1}\{X^T = g(\zeta^T)\}$$

for some surjective $g : \mathscr{Z}^T \to \mathscr{S} \supset \mathscr{V}^T$

$P^*_{X^T,\zeta^T}$:





Jointly Optimal Detector and Watermarking Scheme

Optimization problem:

$$\min_{\gamma,\, P_{X^T,\zeta^T}} \quad MD(\gamma,\, P_{X^T,\zeta} = \mathbb{E}_{P_{X^T,\zeta^T}}[1-\gamma(X^T,\zeta^T)]$$

$$\text{s.t.} \quad \sup_{Q_{X^T}} FA(\gamma, Q_{X^T}, P_{\zeta^T}) \le \alpha \qquad (\Delta_1)$$

$$\mathsf{D}(P_{X^T}, Q_{X^T}) \le \epsilon \qquad (\Delta_2)$$

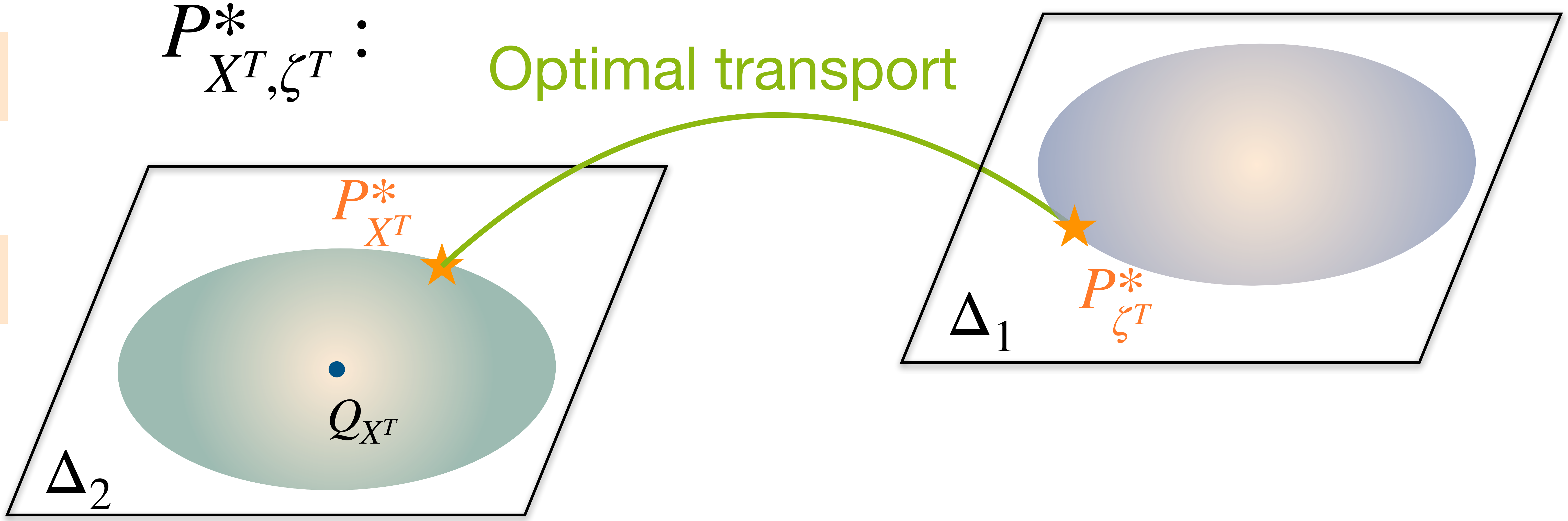
$$P^*_{X^T} = \arg \min_{P_{X^T}:\mathsf{D}(P_{X^T},Q_{X^T}) \le \epsilon} \sum\nolimits_{x^T} (P_{X^T}(x^T) - \alpha)_+$$

✦ **Jointly optimal detector γ^* and watermarking scheme $P^*_{X^T,\zeta^T}$:**

$$\gamma^* = \mathbf{1}\{X^T = g(\zeta^T)\}$$

for some surjective $g : \mathcal{Z}^T \to \mathcal{S} \supset \mathcal{V}^T$

$P^*_{X^T,\zeta^T}$:



Jointly Optimal Detector and Watermarking Scheme

Optimization problem:

$$\min_{\gamma,\, P_{X^T,\zeta^T}} \quad MD(\gamma,\, P_{X^T,\zeta} = \mathbb{E}_{P_{X^T,\zeta^T}}[1-\gamma(X^T,\zeta^T)]$$

s.t. $\sup_{Q_{X^T}} FA(\gamma, Q_{X^T}, P_{\zeta^T}) \leq \alpha$ (Δ_1)

$\mathrm{D}(P_{X^T}, Q_{X^T}) \leq \epsilon$ (Δ_2)

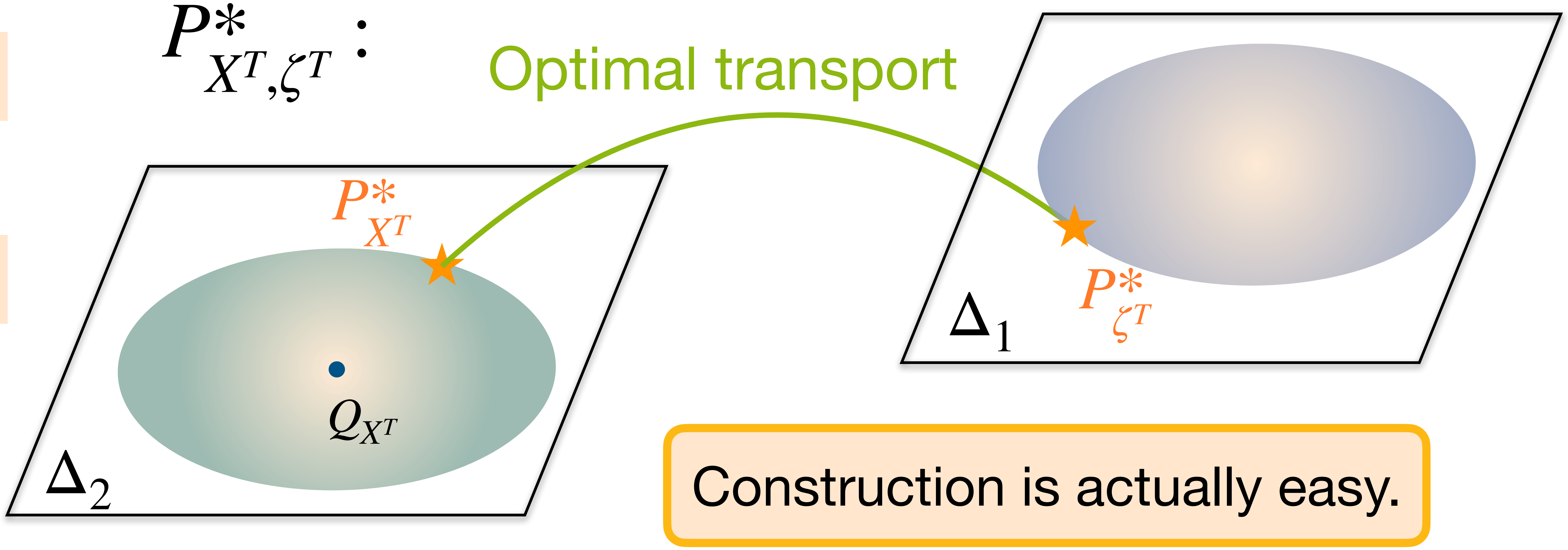
$$P^*_{X^T} = \arg\min_{P_{X^T}:\mathrm{D}(P_{X^T},Q_{X^T})\leq\epsilon} \sum_{x^T} (P_{X^T}(x^T)-\alpha)_+$$

✦ **Jointly optimal detector γ^***
and watermarking scheme $P^*_{X^T,\zeta^T}$:

$$\gamma^* = \mathbf{1}\{X^T = g(\zeta^T)\}$$

for some surjective $g : \mathcal{Z}^T \rightarrow \mathcal{S} \supset \mathcal{V}^T$

$P^*_{X^T,\zeta^T}$:



Construction is actually easy.

Jointly Optimal Detector and Watermarking Scheme

Optimization problem:

$$\min_{\gamma,\, P_{X^T,\zeta^T}} \quad MD(\gamma, P_{X^T,\zeta^T})$$

$$\text{s.t.} \quad \sup_{Q_{X^T}} FA(\gamma, Q_{X^T}, P_{\zeta^T}) \leq \alpha$$

$$\mathrm{D}(P_{X^T}, Q_{X^T}) \leq \epsilon$$

$$P^*_{X^T} = \arg \min_{P_{X^T}: \mathrm{D}(P_{X^T}, Q_{X^T}) \leq \epsilon} \sum_{x^T} (P_{X^T}(x^T) - \alpha)_+$$

✦ **Jointly optimal detector γ^* and watermarking scheme $P^*_{X^T,\zeta^T}$:**

$$\gamma^* = \mathbf{1}\{X^T = g(\zeta^T)\}$$

for some surjective $g : \mathcal{Z}^T \to \mathcal{S} \supset \mathcal{V}^T$

$P^*_{X^T,\zeta^T}$:

Jointly Optimal Detector and Watermarking Scheme

Optimization problem:

$$\min_{\gamma, P_{X^T,\zeta^T}} \quad MD(\gamma, P_{X^T,\zeta^T})$$

s.t. $\sup_{Q_{X^T}} FA(\gamma, Q_{X^T}, P_{\zeta^T}) \le \alpha$

$\mathrm{D}(P_{X^T}, Q_{X^T}) \le \epsilon$

$$P^*_{X^T} = \arg \min_{P_{X^T}:\mathrm{D}(P_{X^T},Q_{X^T}) \le \epsilon} \sum_{x^T} (P_{X^T}(x^T) - \alpha)_+$$

✦ **Jointly optimal detector γ^* and watermarking scheme $P^*_{X^T,\zeta^T}$:**

$$\gamma^* = \mathbf{1}\{X^T = g(\zeta^T)\}$$

for some surjective $g : \mathcal{Z}^T \to \mathcal{S} \supset \mathcal{V}^T$

$P^*_{X^T,\zeta^T}$:

$(T = 1)$

Jointly Optimal Detector and Watermarking Scheme

Optimization problem:

$$\min_{\gamma,\, P_{X^T,\zeta^T}} \quad MD(\gamma, P_{X^T,\zeta^T})$$

$$\text{s.t.} \quad \sup_{Q_{X^T}} FA(\gamma, Q_{X^T}, P_{\zeta^T}) \leq \alpha$$

$$\mathsf{D}(P_{X^T}, Q_{X^T}) \leq \epsilon$$

$$P^*_{X^T} = \arg \min_{P_{X^T}: \mathsf{D}(P_{X^T}, Q_{X^T}) \leq \epsilon} \sum_{x^T} (P_{X^T}(x^T) - \alpha)_+$$

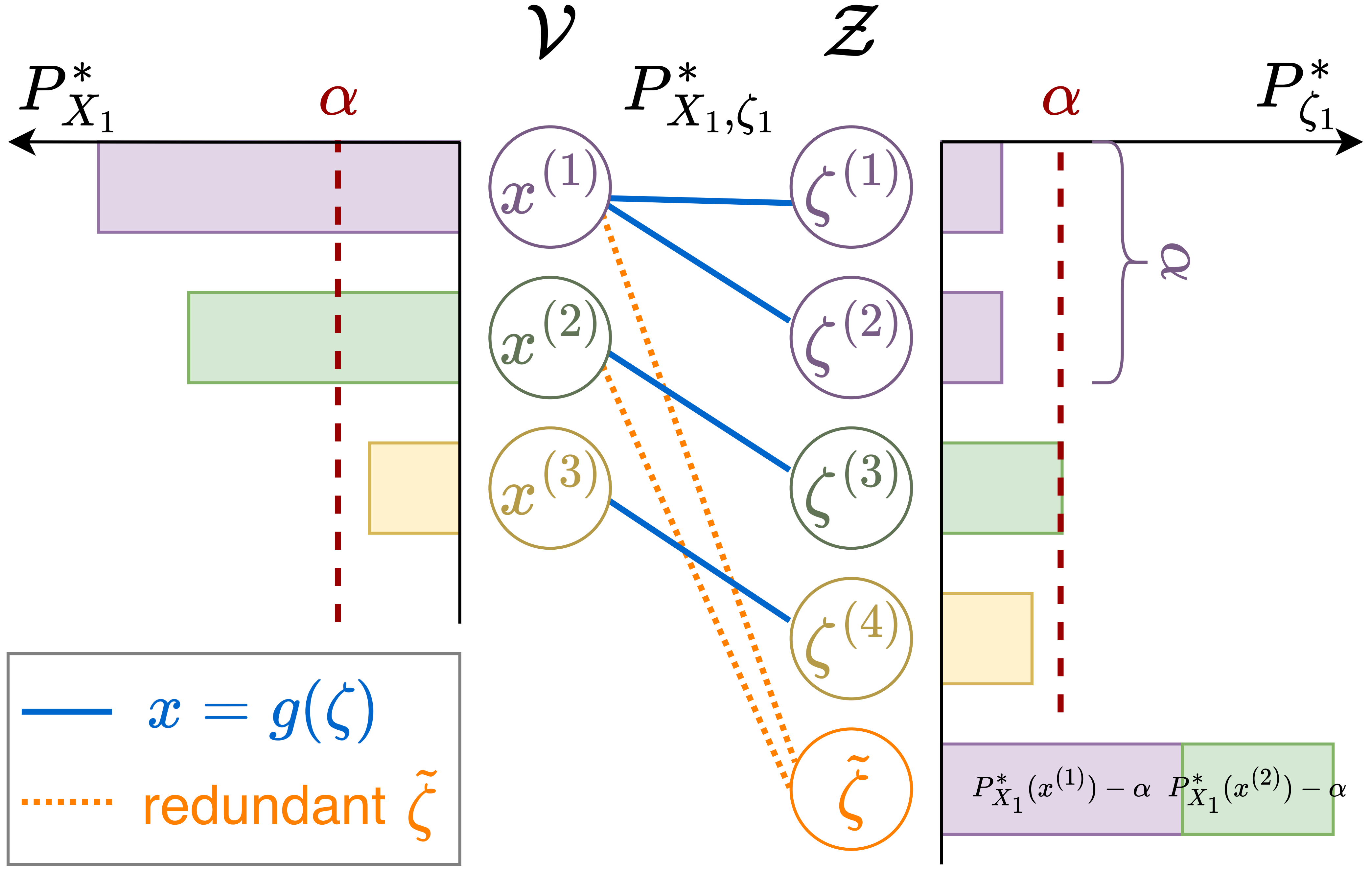
✦ **Jointly optimal detector γ^***
and watermarking scheme $P^*_{X^T,\zeta^T}$:

$$\gamma^* = \mathbf{1}\{X^T = g(\zeta^T)\}$$

for some surjective $g : \mathscr{Z}^T \to \mathcal{S} \supset \mathscr{V}^T$

$P^*_{X^T,\zeta^T}$:

$(T = 1)$



Jointly Optimal Detector and Watermarking Scheme

Optimization problem:

$$\min_{\gamma,\, P_{X^T,\zeta^T}} \quad MD(\gamma,\, P_{X^T,\zeta^T})$$

$$\text{s.t.} \quad \sup_{Q_{X^T}} FA(\gamma, Q_{X^T}, P_{\zeta^T}) \leq \alpha$$

$$\mathrm{D}(P_{X^T}, Q_{X^T}) \leq \epsilon$$

✦ **Jointly optimal detector γ^* and watermarking scheme $P^*_{X^T,\zeta^T}$:**

$$\gamma^* = \mathbf{1}\{X^T = g(\zeta^T)\}$$

for some surjective $g : \mathcal{Z}^T \rightarrow \mathcal{S} \supset \mathcal{V}^T$

$P^*_{X^T,\zeta^T}$:

Jointly Optimal Detector and Watermarking Scheme

Optimization problem:

$$\min_{\gamma,\, P_{X^T,\zeta^T}} \quad MD(\gamma, P_{X^T,\zeta^T})$$

$$\text{s.t.} \quad \sup_{Q_{X^T}} FA(\gamma, Q_{X^T}, P_{\zeta^T}) \le \alpha$$

$$\mathrm{D}(P_{X^T}, Q_{X^T}) \le \epsilon$$

✦ **Jointly optimal detector γ^* and watermarking scheme $P^*_{X^T,\zeta^T}$:**

$$\gamma^* = \mathbf{1}\{X^T = g(\zeta^T)\}$$

for some surjective $g : \mathcal{Z}^T \to \mathcal{S} \supset \mathcal{V}^T$

$P^*_{X^T,\zeta^T}$:

$P^*_{\zeta^T}$ **Adaptive** to original LLM predicted distribution Q_{X^T}

Jointly Optimal Detector and Watermarking Scheme

Optimization problem:

$$\min_{\gamma,\, P_{X^T,\zeta^T}} \quad MD(\gamma, P_{X^T,\zeta^T})$$

s.t. $\sup_{Q_{X^T}} FA(\gamma, Q_{X^T}, P_{\zeta^T}) \leq \alpha$

$\mathrm{D}(P_{X^T}, Q_{X^T}) \leq \epsilon$

✦ **Jointly optimal detector γ^* and watermarking scheme $P^*_{X^T,\zeta^T}$:**

$$\gamma^* = \mathbf{1}\{X^T = g(\zeta^T)\}$$

for some surjective $g : \mathcal{Z}^T \to \mathcal{S} \supset \mathcal{V}^T$

$P^*_{X^T,\zeta^T}$:

$P^*_{\zeta^T}$ **Adaptive** to original LLM predicted distribution Q_{X^T}

↓

Unlike existing watermarking methods

Sequence-Level Optimal to Token-Level Optimal

Sequence-Level Optimal to Token-Level Optimal

- Previous optimal result holds for **fixed T** ⇒ unable to implement dynamically

Sequence-Level Optimal to Token-Level Optimal

- Previous optimal result holds for **fixed** $T \Rightarrow$ unable to implement dynamically
- **Solution:** implement it token by token

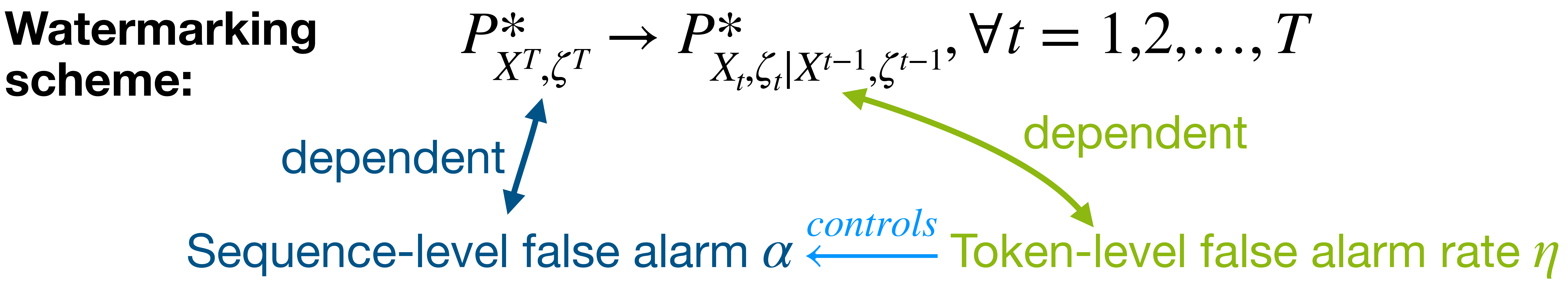
Sequence-Level Optimal to Token-Level Optimal

- Previous optimal result holds for **fixed** $T \Rightarrow$ unable to implement dynamically
- **Solution:** implement it token by token

Watermarking scheme: $P^*_{X^T,\zeta^T} \to P^*_{X_t,\zeta_t|X^{t-1},\zeta^{t-1}}, \forall t = 1,2,\ldots,T$

Sequence-Level Optimal to Token-Level Optimal

- Previous optimal result holds for **fixed** $T \Rightarrow$ unable to implement dynamically
- **Solution:** implement it token by token



Watermarking scheme: $P^*_{X^T,\zeta^T} \to P^*_{X_t,\zeta_t|X^{t-1},\zeta^{t-1}}, \forall t = 1,2,\ldots,T$

dependent

dependent

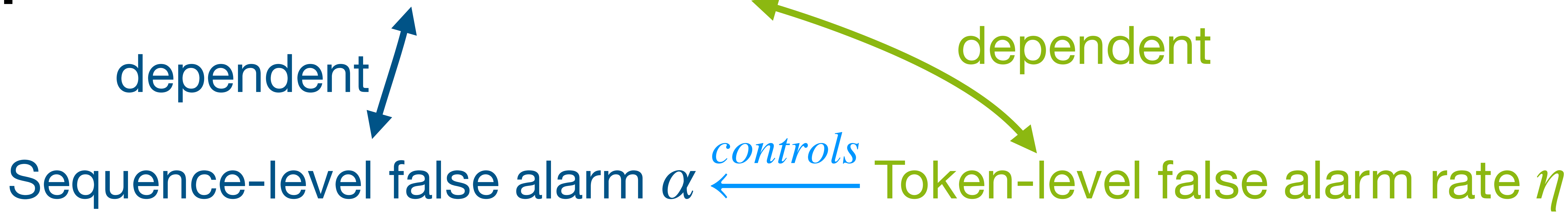
Sequence-level false alarm α $\xleftarrow{\textit{controls}}$ Token-level false alarm rate η

Sequence-Level Optimal to Token-Level Optimal

- Previous optimal result holds for **fixed** $T \Rightarrow$ unable to implement dynamically
- **Solution:** implement it token by token

Watermarking scheme:

$$P^*_{X^T,\zeta^T} \to P^*_{X_t,\zeta_t|X^{t-1},\zeta^{t-1}}, \forall t = 1,2,\ldots,T$$



dependent

dependent

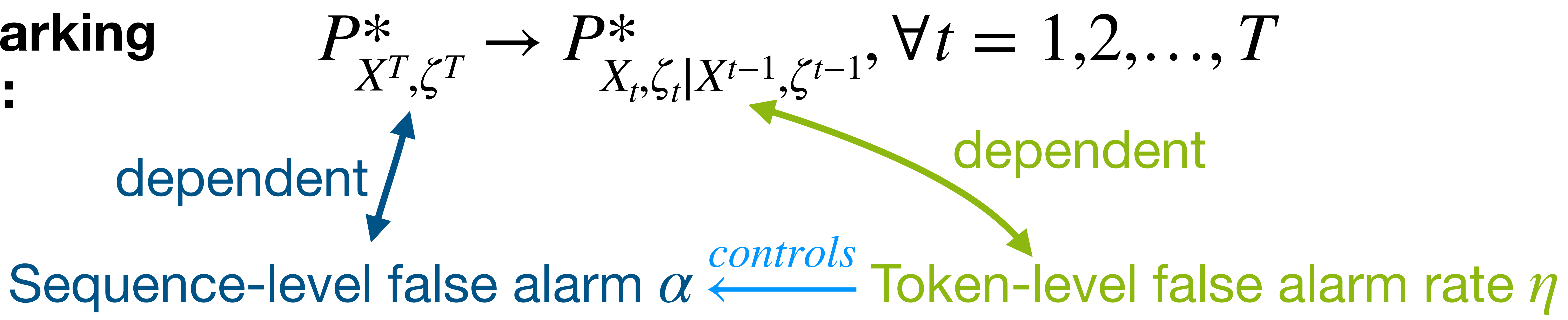
Sequence-level false alarm $\alpha \xleftarrow{\textit{controls}}$ Token-level false alarm rate η

Detector:

Sequence-Level Optimal to Token-Level Optimal

- Previous optimal result holds for **fixed** T ⇒ unable to implement dynamically
- **Solution:** implement it token by token

Watermarking scheme:



$$P^*_{X^T,\zeta^T} \to P^*_{X_t,\zeta_t|X^{t-1},\zeta^{t-1}}, \forall t = 1,2,\ldots,T$$

dependent

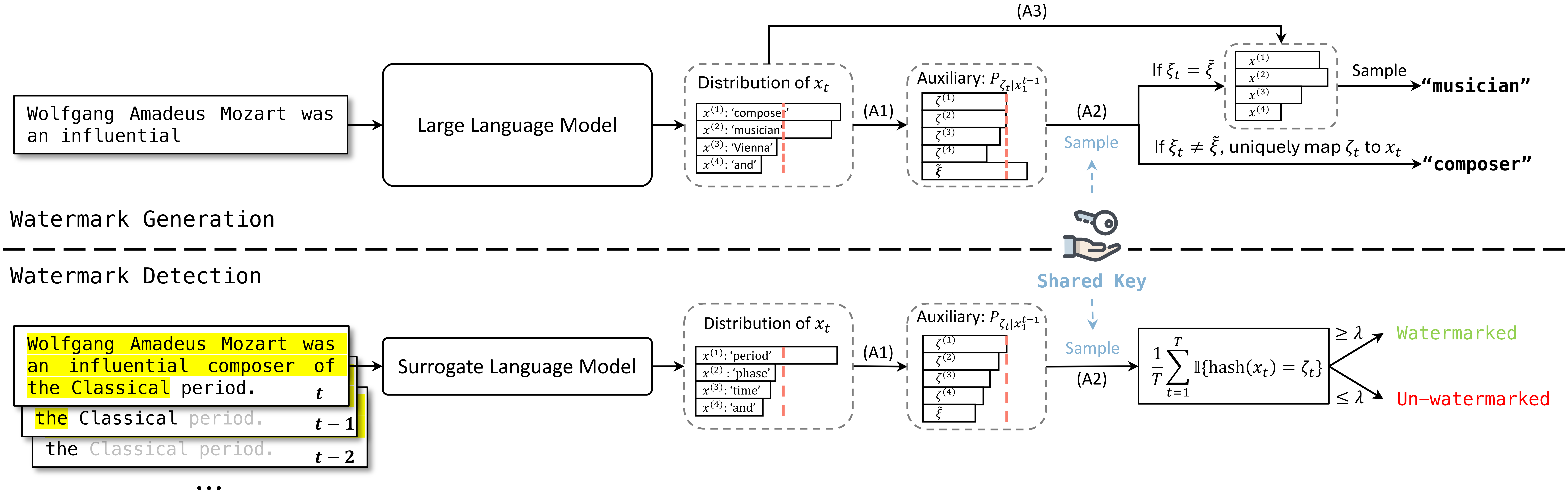
dependent

Sequence-level false alarm α $\xleftarrow{\textit{controls}}$ Token-level false alarm rate η

Detector:

$$\gamma_{\text{tk}} = \mathbf{1}\left\{\frac{1}{T}\sum_{t=1}^{T} \mathbf{1}\{X_t = g(\zeta_t)\} \ge \lambda\right\}$$ for some surjective $g : \mathscr{Z} \to \mathscr{S} \supset \mathscr{V}$

DAWA: Distribution-Adaptive Watermarking Algorithm ($\epsilon = 0$, distortion-free)



(A3)

Wolfgang Amadeus Mozart was an influential

Large Language Model

Distribution of x_t

$x^{(1)}$: 'composer'

$x^{(2)}$: 'musician'

$x^{(3)}$: 'Vienna'

$x^{(4)}$: 'and'

(A1)

Auxiliary: $P_{\zeta_t|x_1^{t-1}}$

$\zeta^{(1)}$

$\zeta^{(2)}$

$\zeta^{(3)}$

$\zeta^{(4)}$

$\tilde{\xi}$

(A2)

Sample

If $\xi_t = \tilde{\xi}$

$x^{(1)}$

$x^{(2)}$

$x^{(3)}$

$x^{(4)}$

Sample

"musician"

If $\xi_t \neq \tilde{\xi}$, uniquely map ζ_t to x_t

"composer"

Watermark Generation

Shared Key

Watermark Detection

Wolfgang Amadeus Mozart was an influential composer of the Classical period. t

the Classical period. $t-1$

the Classical period. $t-2$

...

Surrogate Language Model

Distribution of x_t

$x^{(1)}$: 'period'

$x^{(2)}$: 'phase'

$x^{(3)}$: 'time'

$x^{(4)}$: 'and'

(A1)

Auxiliary: $P_{\zeta_t|x_1^{t-1}}$

$\zeta^{(1)}$

$\zeta^{(2)}$

$\zeta^{(3)}$

$\zeta^{(4)}$

$\tilde{\xi}$

Sample

(A2)

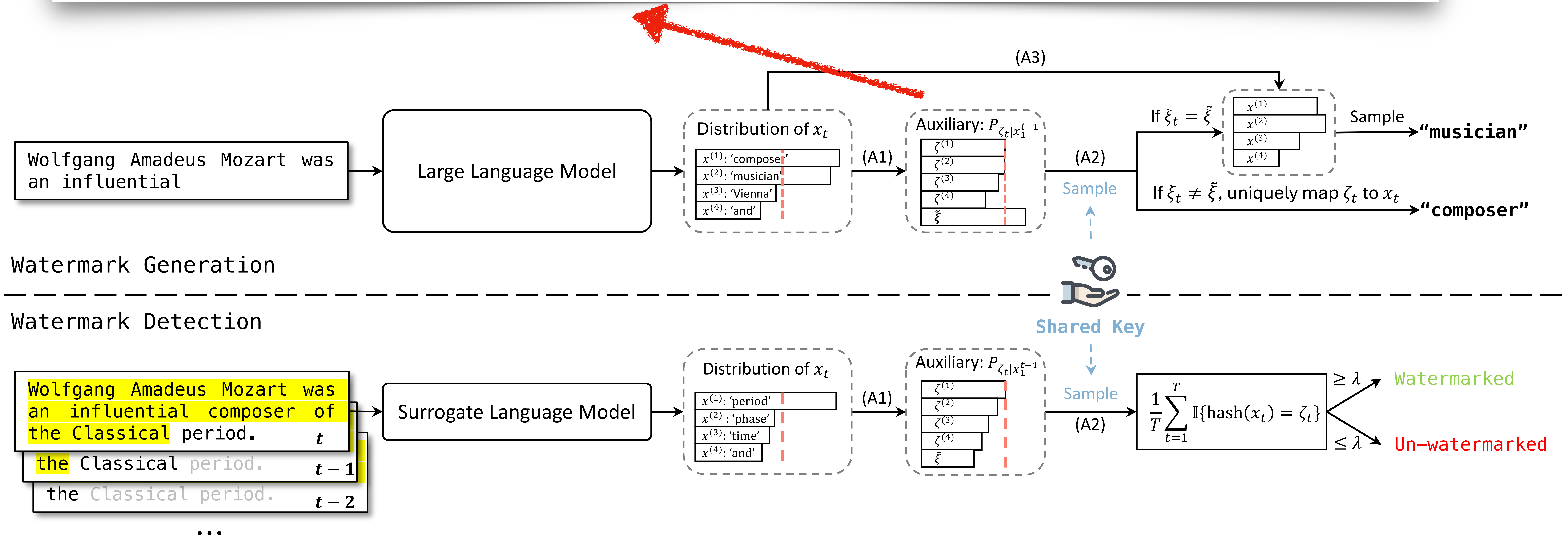
$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{I}\{\text{hash}(x_t) = \zeta_t\}$$

$\geq \lambda$ Watermarked

$\leq \lambda$ Un-watermarked

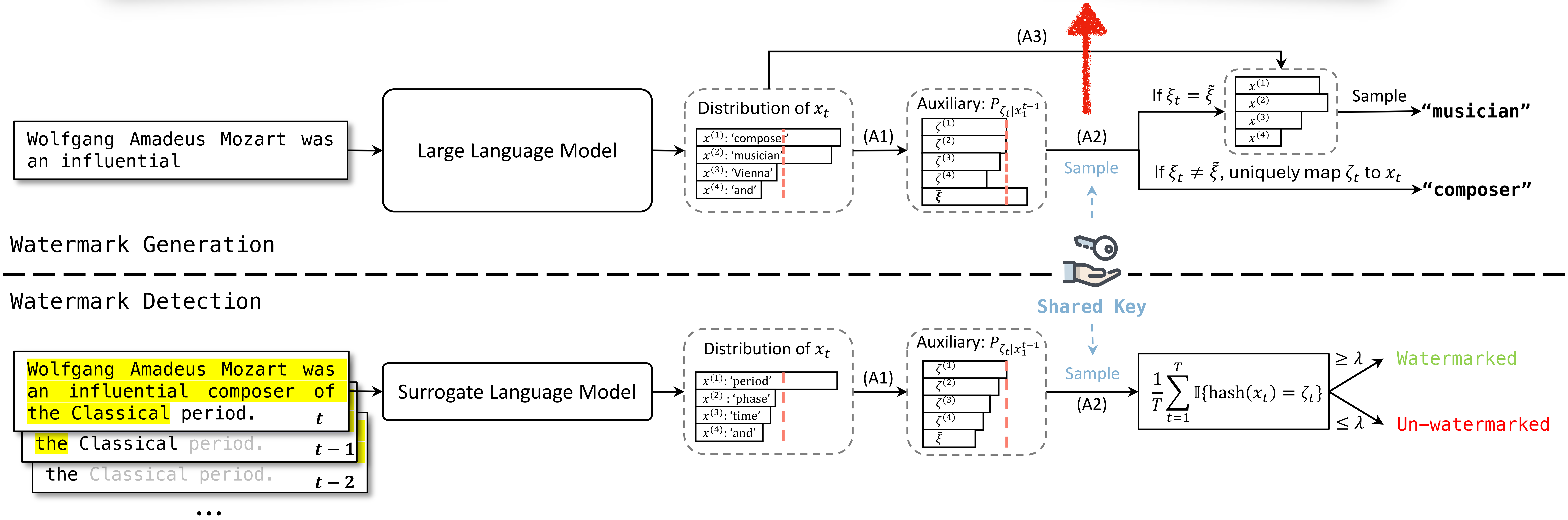
DAWA: Distribution-Adaptive Watermarking Algorithm ($\epsilon = 0$, distortion-free)

At each time t, construct $P^*_{\zeta_t|X_1^t}$ from the LLM predicted distribution $Q_{X_t|X_1^{t-1}}$



DAWA: Distribution-Adaptive Watermarking Algorithm ($\epsilon = 0$, distortion-free)

Sample ζ_t using Gumbel max trick: $\zeta_t \leftarrow \arg\max_{\zeta} \log P^*_{\zeta_t|x_1^t}(\zeta) + G_{\zeta,t}$



(A3)

Wolfgang Amadeus Mozart was an influential

Large Language Model

Distribution of x_t

$x^{(1)}$: 'composer'

$x^{(2)}$: 'musician'

$x^{(3)}$: 'Vienna'

$x^{(4)}$: 'and'

(A1)

Auxiliary: $P_{\zeta_t|x_1^{t-1}}$

$\zeta^{(1)}$

$\zeta^{(2)}$

$\zeta^{(3)}$

$\zeta^{(4)}$

$\tilde{\xi}$

(A2)

Sample

If $\xi_t = \tilde{\xi}$

$x^{(1)}$

$x^{(2)}$

$x^{(3)}$

$x^{(4)}$

Sample

"musician"

If $\xi_t \neq \tilde{\xi}$, uniquely map ζ_t to x_t

"composer"

Watermark Generation

Shared Key

Watermark Detection

Wolfgang Amadeus Mozart was an influential composer of the Classical period. t

the Classical period. $t-1$

the Classical period. $t-2$

...

Surrogate Language Model

Distribution of x_t

$x^{(1)}$: 'period'

$x^{(2)}$: 'phase'

$x^{(3)}$: 'time'

$x^{(4)}$: 'and'

(A1)

Auxiliary: $P_{\zeta_t|x_1^{t-1}}$

$\zeta^{(1)}$

$\zeta^{(2)}$

$\zeta^{(3)}$

$\zeta^{(4)}$

$\tilde{\xi}$

Sample

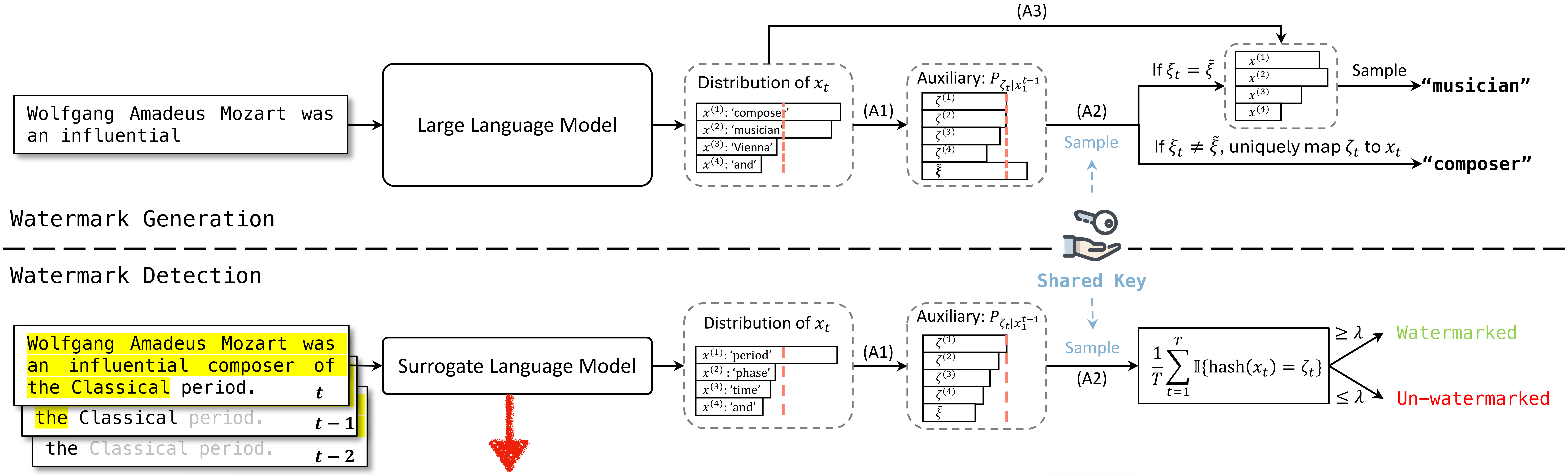
(A2)

$\frac{1}{T}\sum_{t=1}^{T} \mathbb{I}\{\text{hash}(x_t) = \zeta_t\}$

$\geq \lambda$ Watermarked

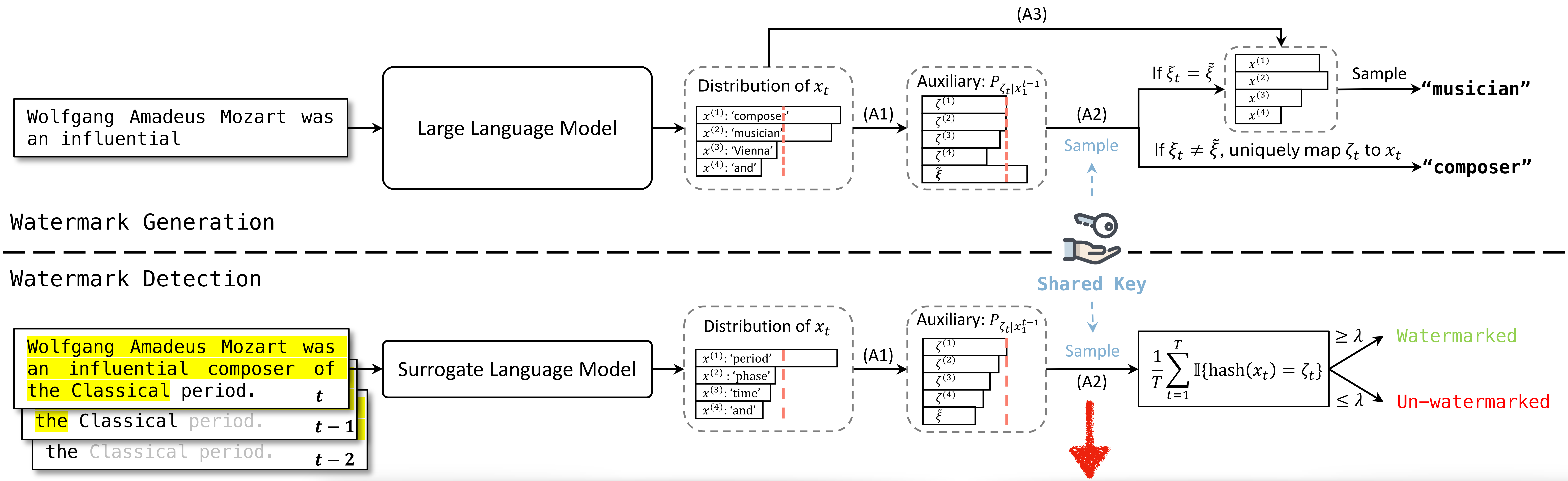
$\leq \lambda$ Un-watermarked

DAWA: Distribution-Adaptive Watermarking Algorithm ($\epsilon = 0$, distortion-free)



Approximate distribution of X_t so as to construct $\tilde{P}_{\zeta_t|x_1^t}$

DAWA: Distribution-Adaptive Watermarking Algorithm ($\epsilon = 0$, distortion-free)



…

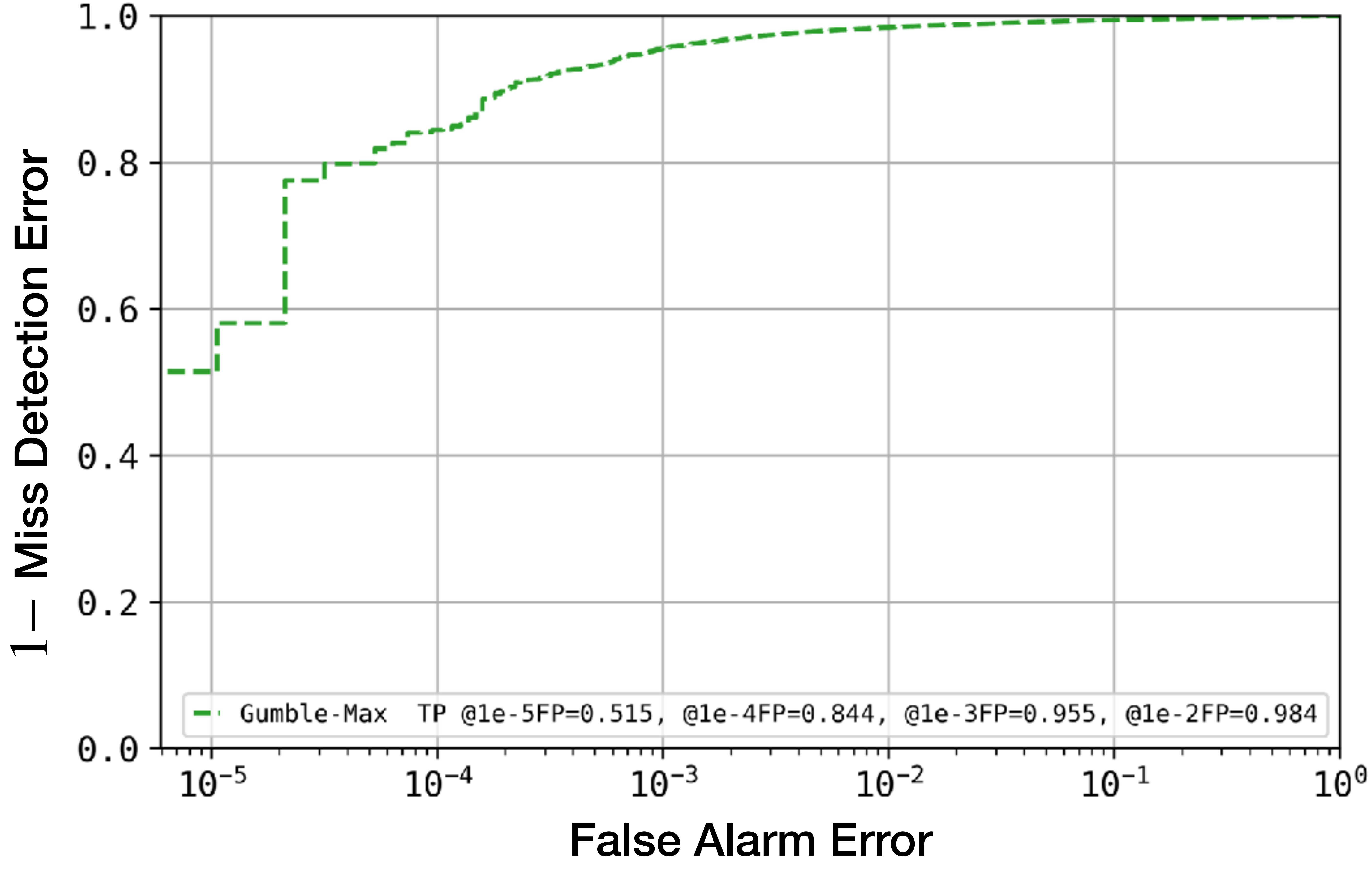
Sample ζ_t using Gumbel max trick: $\zeta_t \leftarrow \arg\max_{\zeta} \log \tilde{P}_{\zeta_t|x_1^t}(\zeta) + G_{\zeta,t}$

From Theory to Practical Algorithm

DAWA (**D**istribution-**A**daptive **W**atermarking **A**lgorithm)

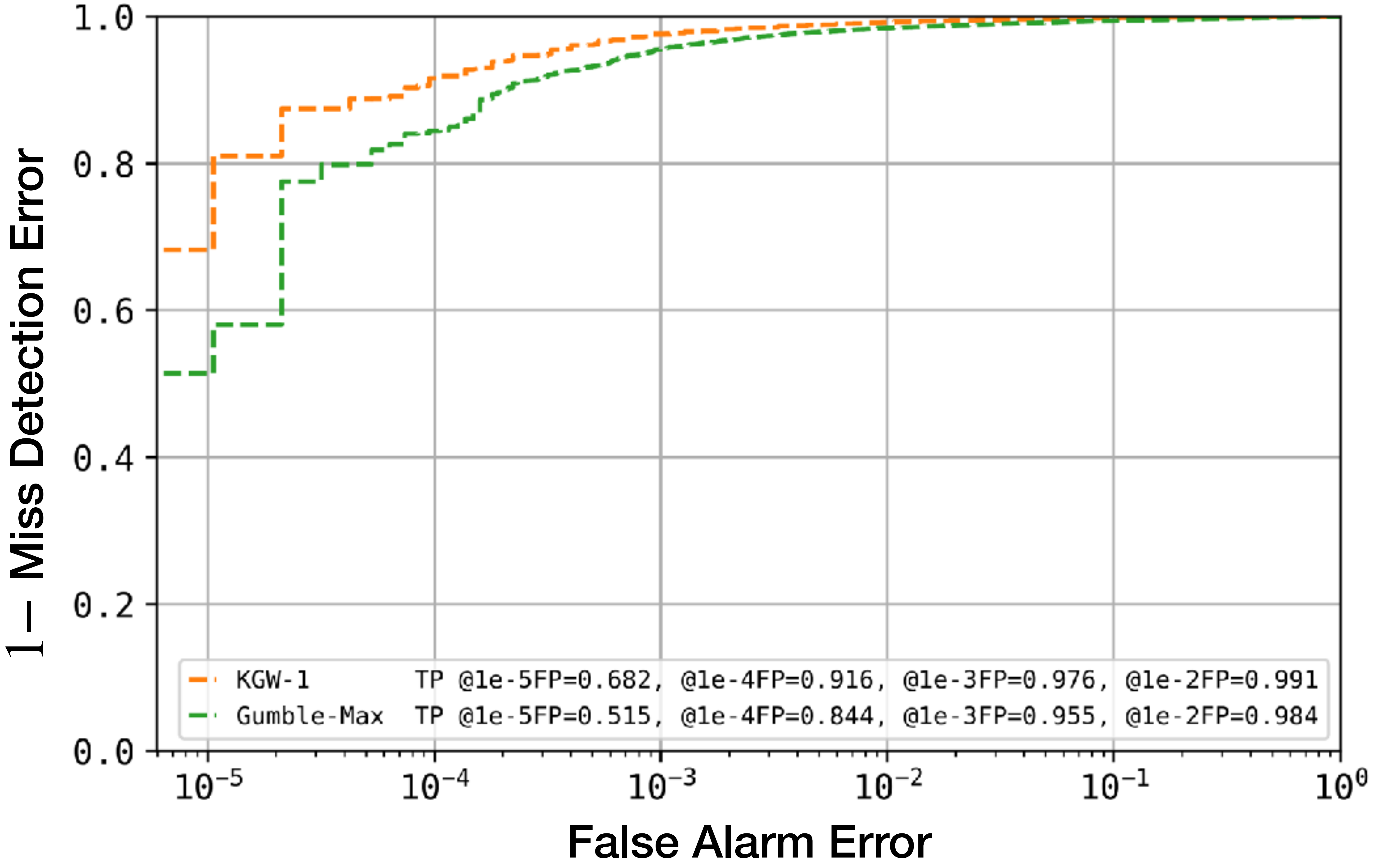
From Theory to Practical Algorithm

DAWA (**D**istribution-**A**daptive **W**atermarking **A**lgorithm)



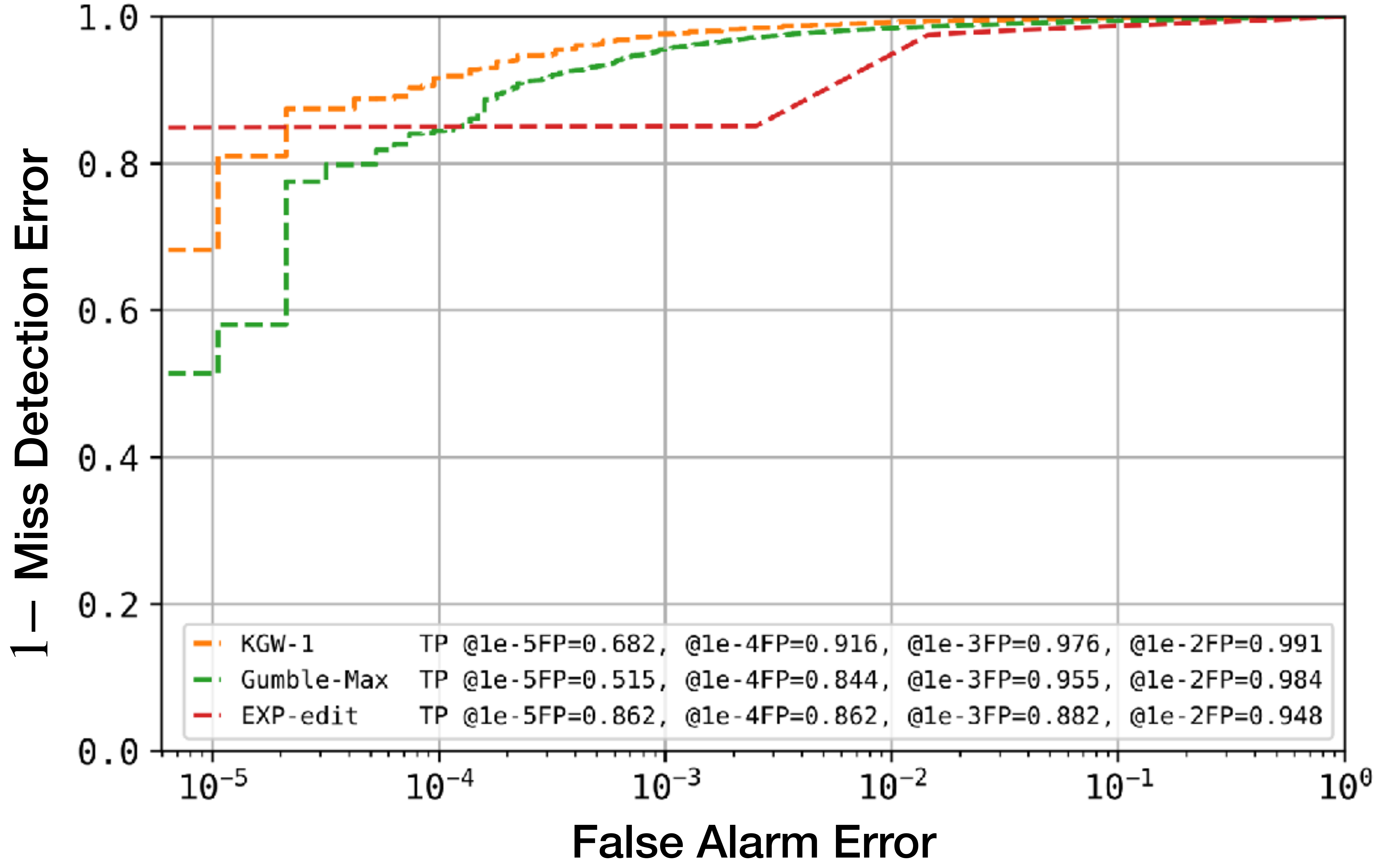
From Theory to Practical Algorithm

DAWA (**D**istribution-**A**daptive **W**atermarking **A**lgorithm)



From Theory to Practical Algorithm

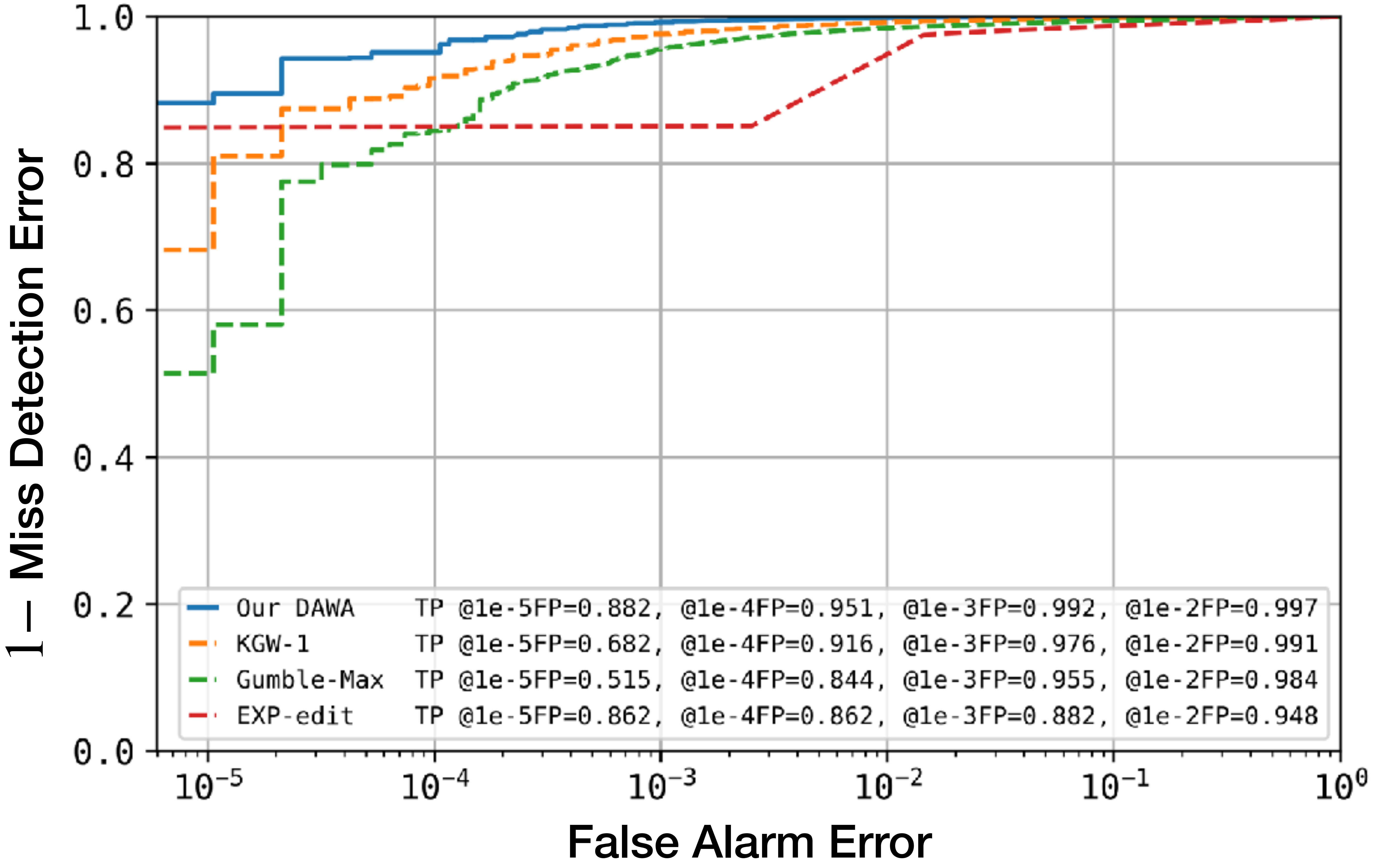
DAWA (**D**istribution-**A**daptive **W**atermarking **A**lgorithm)



From Theory to Practical Algorithm

DAWA (**D**istribution-**A**daptive **W**atermarking **A**lgorithm)

Fast and Accurate

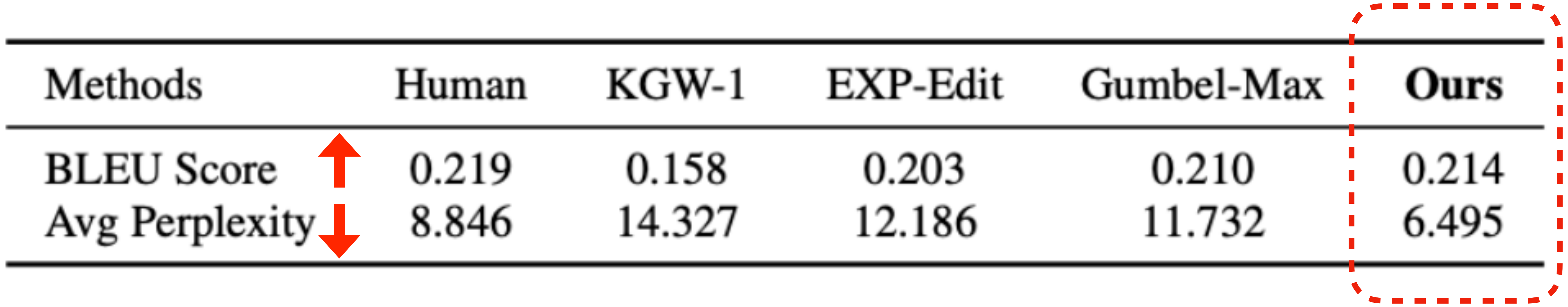


From Theory to Practical Algorithm

DAWA (**D**istribution-**A**daptive **W**atermarking **A**lgorithm)

Fast and Accurate

Text quality high



Methods	Human	KGW-1	EXP-Edit	Gumbel-Max	**Ours**
BLEU Score ↑	0.219	0.158	0.203	0.210	0.214
Avg Perplexity ↓	8.846	14.327	12.186	11.732	6.495

With Text Modifications?

Original Text x^T

We propose a pipeline to inject multi-bit text watermark. We encode the watermark by paraphrasing a piece of text using special paraphrasers. Then the watermark can be detected by our trained decoder.

$\tilde{x}^T$ **E.g. Paraphrased Text**

We propose a method for multi-bit text watermark injection. The watermark is encoded into a piece of text with special paraphrasers. We then detect the watermark using our trained decoder.

With Text Modifications?

Original Text x^T

We propose a pipeline to inject multi-bit text watermark. We encode the watermark by paraphrasing a piece of text using special paraphrasers. Then the watermark can be detected by our trained decoder.

$\tilde{x}^T$ **E.g. Paraphrased Text**

We propose a method for multi-bit text watermark injection. The watermark is encoded into a piece of text with special paraphrasers. We then detect the watermark using our trained decoder.

$x^T \neq \tilde{x}^T$ **but the same meaning** $h(x^T) = h(\tilde{x}^T)$

With Text Modifications?

Original Text x^T

We propose a pipeline to inject multi-bit text watermark. We encode the watermark by paraphrasing a piece of text using special paraphrasers. Then the watermark can be detected by our trained decoder.

→

$\tilde{x}^T$ **E.g. Paraphrased Text**

We propose a method for multi-bit text watermark injection. The watermark is encoded into a piece of text with special paraphrasers. We then detect the watermark using our trained decoder.

$x^T \neq \tilde{x}^T$ **but the same meaning** $h(x^T) = h(\tilde{x}^T)$

- $h : \mathscr{V}^T \to [K]$: maps x^T to a finite latent space $[K]$, e.g., a semantic mapping

With Text Modifications?

Original Text x^T

We propose a pipeline to inject multi-bit text watermark. We encode the watermark by paraphrasing a piece of text using special paraphrasers. Then the watermark can be detected by our trained decoder.

→

$\tilde{x}^T$ **E.g. Paraphrased Text**

We propose a method for multi-bit text watermark injection. The watermark is encoded into a piece of text with special paraphrasers. We then detect the watermark using our trained decoder.

$x^T \neq \tilde{x}^T$ **but the same meaning** $h(x^T) = h(\tilde{x}^T)$

- $h : \mathscr{V}^T \rightarrow [K]$: maps x^T to a finite latent space $[K]$, e.g., a semantic mapping
- Equivalent class: $\mathscr{B}_h(x^T) = \{\tilde{x}^T \in \mathscr{V}^T : h(\tilde{x}^T) = h(x^T)\}$

Performance Metric with Text Modifications

Performance Metric with Text Modifications

- **Text modification:** x^T can be modified as any text within $\mathscr{B}_h(x^T)$

Performance Metric with Text Modifications

- **Text modification:** x^T can be modified as any text within $\mathscr{B}_h(x^T)$
- h-robust Type-I and Type-II errors:

Performance Metric with Text Modifications

- **Text modification:** x^T can be modified as any text within $\mathscr{B}_h(x^T)$
- h-robust Type-I and Type-II errors:

$$FA(\gamma, Q_{X^T}, P_{\zeta^T}, h) := \mathbb{E}_{Q_{X^T}\otimes P_{\zeta^T}}\left[\sup_{\tilde{x}^T\in\mathscr{B}_h(X^T)} \mathbf{1}\{\gamma(\tilde{x}^T, \zeta^T) = 1\}\right]$$

Performance Metric with Text Modifications

- **Text modification:** x^T can be modified as any text within $\mathscr{B}_h(x^T)$
- h-robust Type-I and Type-II errors:

$$FA(\gamma, Q_{X^T}, P_{\zeta^T}, h) := \mathbb{E}_{Q_{X^T} \otimes P_{\zeta^T}} \left[\sup_{\tilde{x}^T \in \mathscr{B}_h(X^T)} \mathbf{1}\{\gamma(\tilde{x}^T, \zeta^T) = 1\} \right]$$

$$MD(\gamma, P_{X^T, \zeta^T}, h) := \mathbb{E}_{P_{X^T, \zeta^T}} \left[\sup_{\tilde{x}^T \in \mathscr{B}_h(X^T)} \mathbf{1}\{\gamma(\tilde{x}^T, \zeta^T) = 0\} \right]$$

Watermarking Robust Against Text Modifications

Optimization problem:

$$\min_{\gamma,\, P_{X^T,\zeta^T}} \quad MD(\gamma,\, P_{X^T,\zeta^T},\, h)$$

$$\text{s.t.} \quad \sup_{Q_{X^T}} FA(\gamma, Q_{X^T}, P_{\zeta^T}, h) \leq \alpha$$

$$\mathrm{D}(P_{X^T}, Q_{X^T}) \leq \epsilon$$

Watermarking Robust Against Text Modifications

Optimization problem:

$$\min_{\gamma,\, P_{X^T,\zeta^T}} \quad MD(\gamma, P_{X^T,\zeta^T}, h)$$

$$\text{s.t.} \quad \sup_{Q_{X^T}} FA(\gamma, Q_{X^T}, P_{\zeta^T}, h) \leq \alpha$$

$$\mathrm{D}(P_{X^T}, Q_{X^T}) \leq \epsilon$$

✦ **Minimum h-robust miss-detection error:**

Watermarking Robust Against Text Modifications

Optimization problem:

$$\min_{\gamma,\, P_{X^T,\zeta^T}} \quad MD(\gamma,\, P_{X^T,\zeta^T},\, h)$$

$$\text{s.t.} \quad \sup_{Q_{X^T}} FA(\gamma, Q_{X^T}, P_{\zeta^T},\, h) \leq \alpha$$

$$\mathrm{D}(P_{X^T}, Q_{X^T}) \leq \epsilon$$

✦ **Minimum h-robust miss-detection error:**

$$\beta_1^*(Q_{X^T}, \alpha, \epsilon, h)$$

$$= \min_{P_{X^T}:\mathrm{D}(P_{X^T},Q_{X^T})\leq\epsilon} \sum_{k\in[K]} \left(\left(\sum_{x^T:h(x^T)=k} P_{X^T}(x^T) \right) - \alpha \right)_+$$

Watermarking Robust Against Text Modifications

Optimization problem:

$$\min_{\gamma,\, P_{X^T,\zeta^T}} \quad MD(\gamma,\, P_{X^T,\zeta^T},\, h)$$

$$\text{s.t.} \quad \sup_{Q_{X^T}} FA(\gamma, Q_{X^T}, P_{\zeta^T},\, h) \le \alpha$$

$$\mathsf{D}(P_{X^T}, Q_{X^T}) \le \epsilon$$

✦ **Minimum h-robust miss-detection error:**

$$\beta_1^*(Q_{X^T}, \alpha, \epsilon, h)$$

$$= \min_{P_{X^T}:\mathsf{D}(P_{X^T},Q_{X^T})\le\epsilon} \sum_{k\in[K]} \left(\left(\sum_{x^T:h(x^T)=k} P_{X^T}(x^T)\right) - \alpha\right)_+$$

Higher than the minimum miss-detection error without considering robustness

Watermarking Robust Against Text Modifications

Optimization problem:

$$\min_{\gamma,\, P_{X^T,\zeta^T}} \quad MD(\gamma,\, P_{X^T,\zeta^T},\, h)$$

$$\text{s.t.} \quad \sup_{Q_{X^T}} FA(\gamma, Q_{X^T}, P_{\zeta^T},\, h) \le \alpha$$

$$\mathrm{D}(P_{X^T}, Q_{X^T}) \le \epsilon$$

✦ **Minimum h-robust miss-detection error:**

$$\beta_1^*(Q_{X^T}, \alpha, \epsilon, h)$$

$$= \min_{P_{X^T}:\mathrm{D}(P_{X^T},Q_{X^T})\le\epsilon} \sum_{k\in[K]} \left(\left(\sum_{x^T:h(x^T)=k} P_{X^T}(x^T)\right) - \alpha\right)_+$$

Higher than the minimum miss-detection error without considering robustness

✦ **Optimal watermarking scheme:**

add signal ζ^T to $P_{h(X^T)}$, e.g., in the semantic space

Watermarking Robust Against Text Modifications

Optimization problem:

$$\min_{\gamma,\, P_{X^T,\zeta^T}} \quad MD(\gamma,\, P_{X^T,\zeta^T},\, h)$$

$$\text{s.t.} \quad \sup_{Q_{X^T}} FA(\gamma, Q_{X^T}, P_{\zeta^T},\, h) \le \alpha$$

$$\mathsf{D}(P_{X^T}, Q_{X^T}) \le \epsilon$$

✦ **Minimum h-robust miss-detection error:**

$$\beta_1^*(Q_{X^T}, \alpha, \epsilon, h)$$

$$= \min_{P_{X^T}:\mathsf{D}(P_{X^T},Q_{X^T})\le\epsilon} \sum_{k\in[K]} \left(\left(\sum_{x^T:h(x^T)=k} P_{X^T}(x^T) \right) - \alpha \right)_+$$

Higher than the minimum miss-detection error without considering robustness

✦ **Optimal watermarking scheme:**

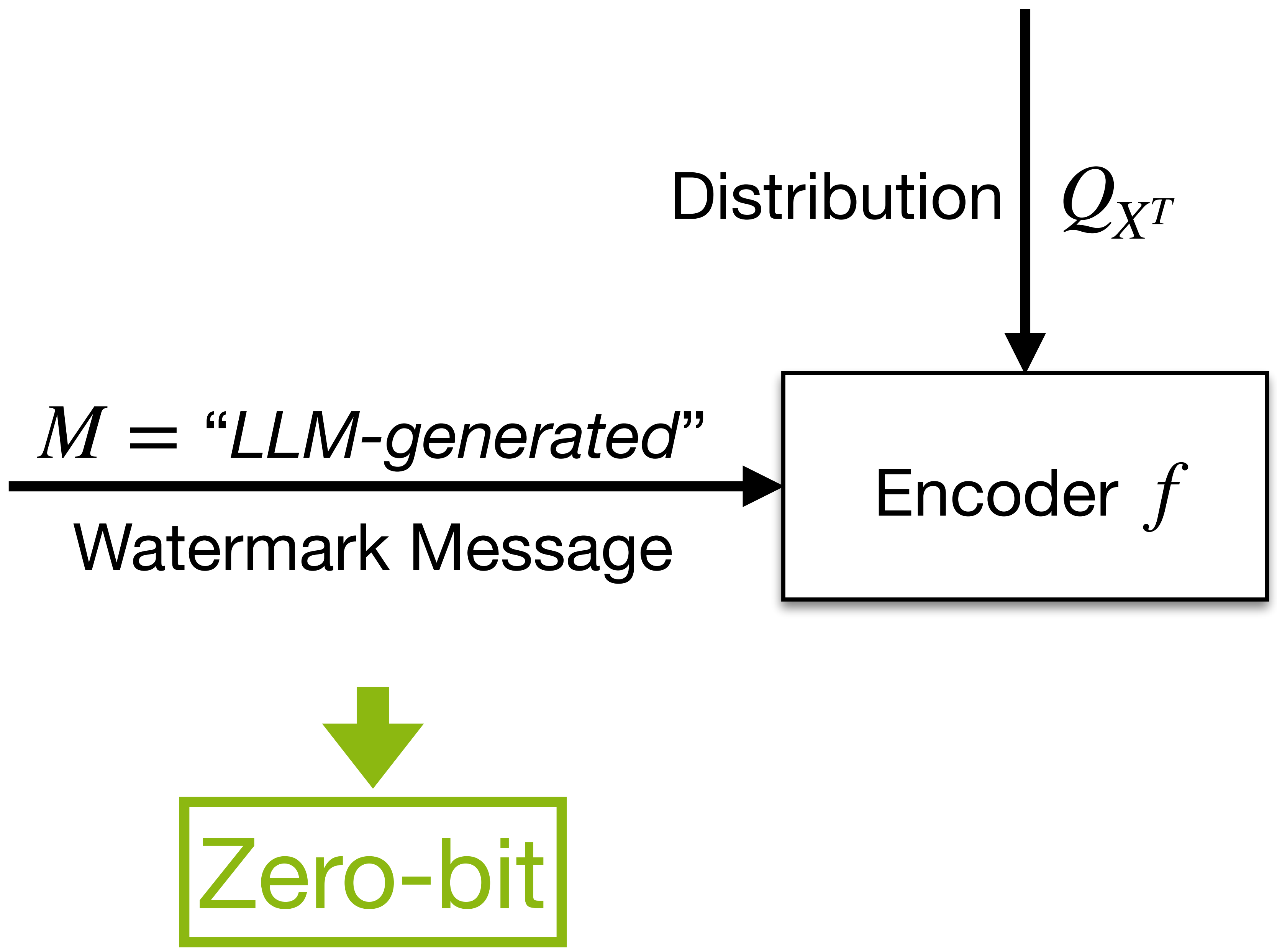
Future work

add signal ζ^T to $P_{h(X^T)}$, e.g., in the semantic space

Want to embed more watermark message?

e.g. LLM ID, User ID, Content Summary…

Distributional Information Embedding with Side Information——Multi-bit Watermarking



Distribution Q_{X^T}

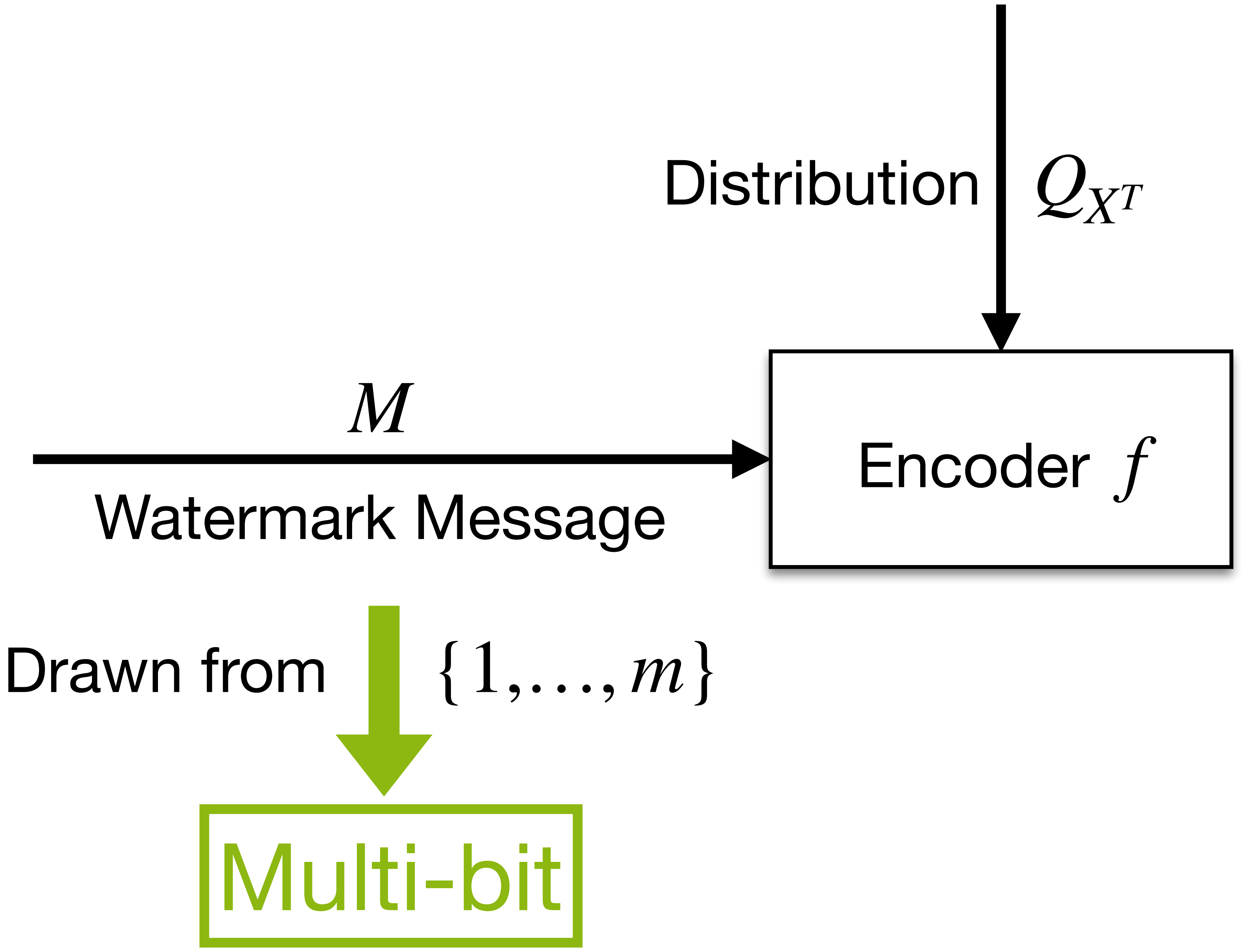
$M =$ "*LLM-generated*"

Watermark Message

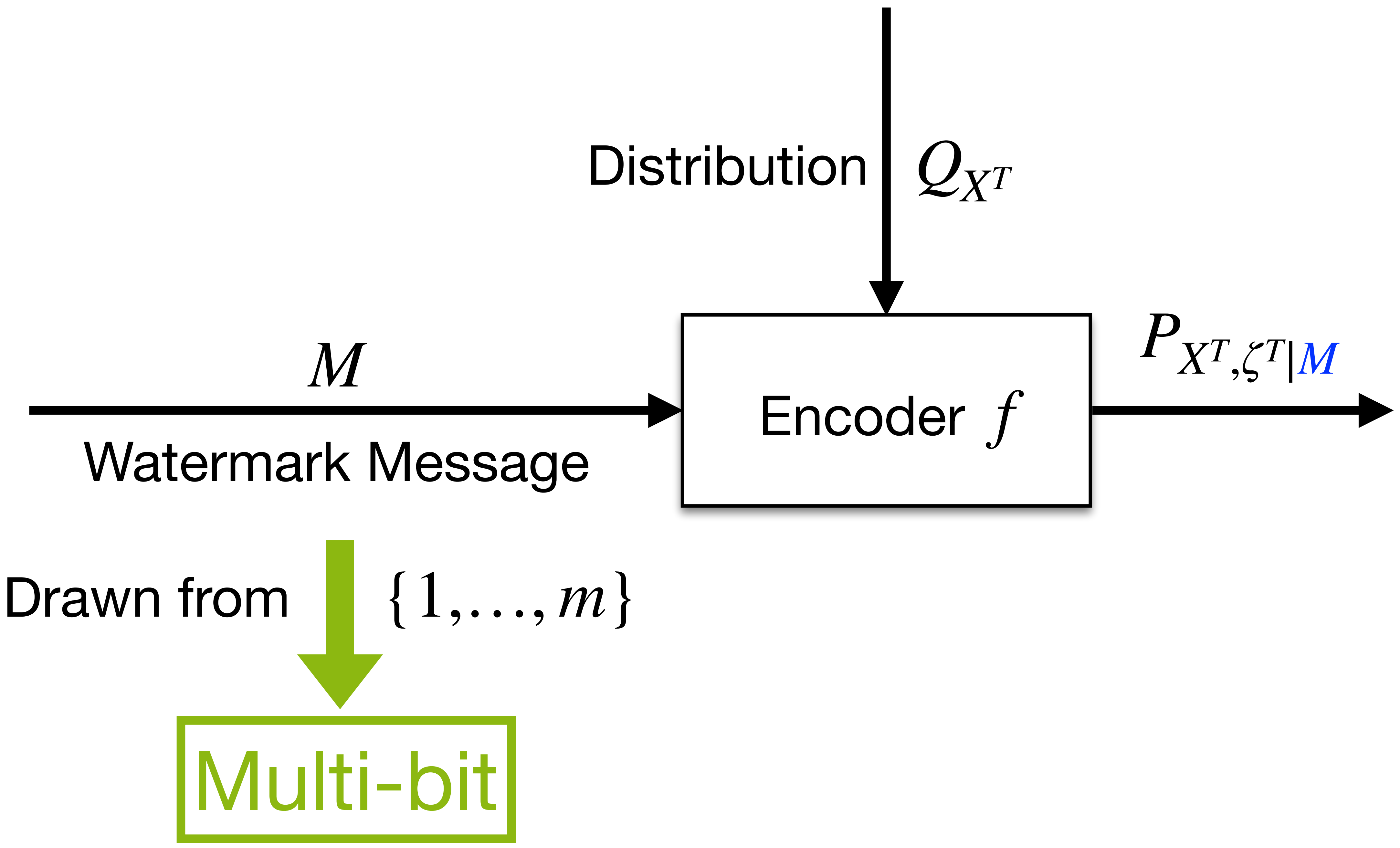
Encoder f

Zero-bit

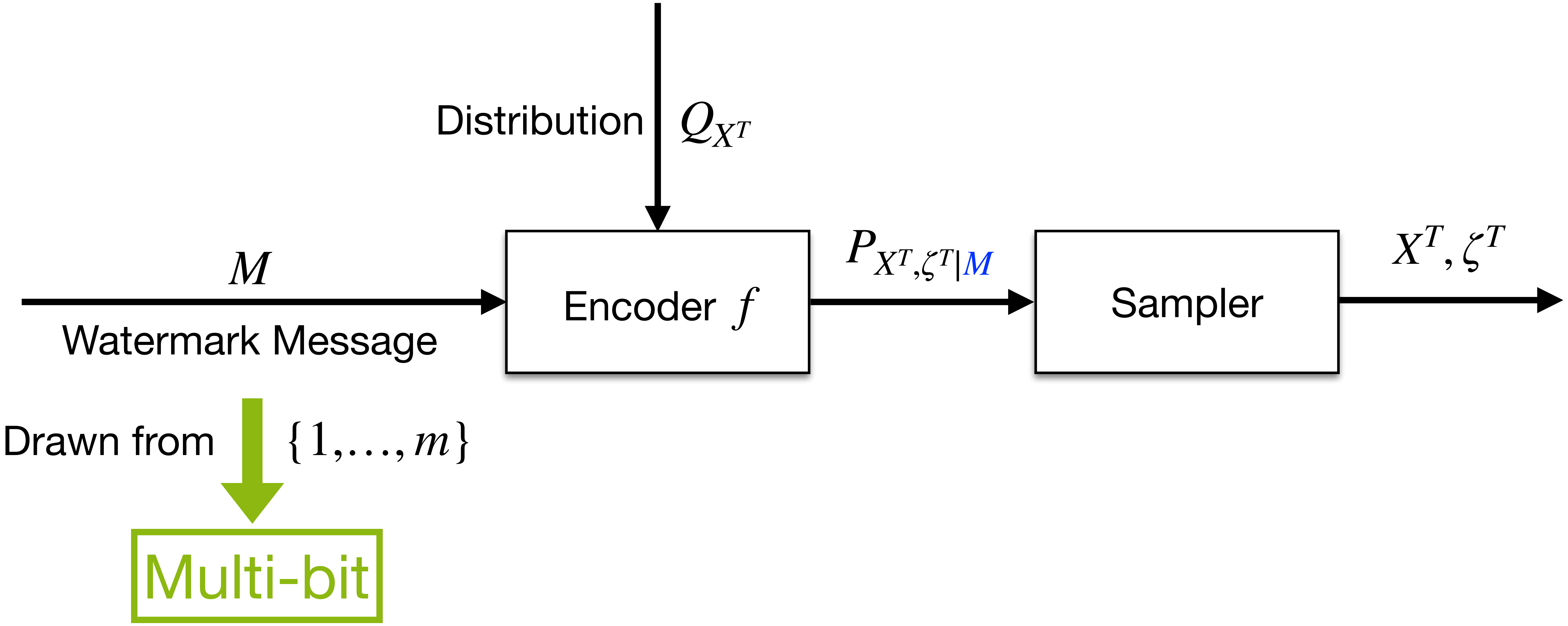
Distributional Information Embedding with Side Information——Multi-bit Watermarking



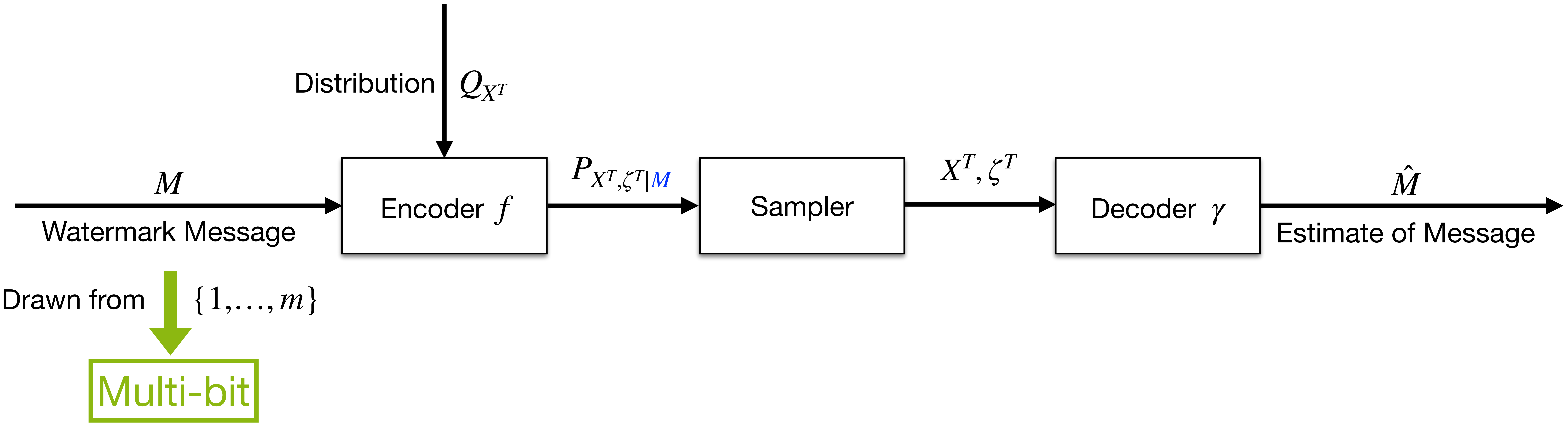
Distributional Information Embedding with Side Information——Multi-bit Watermarking



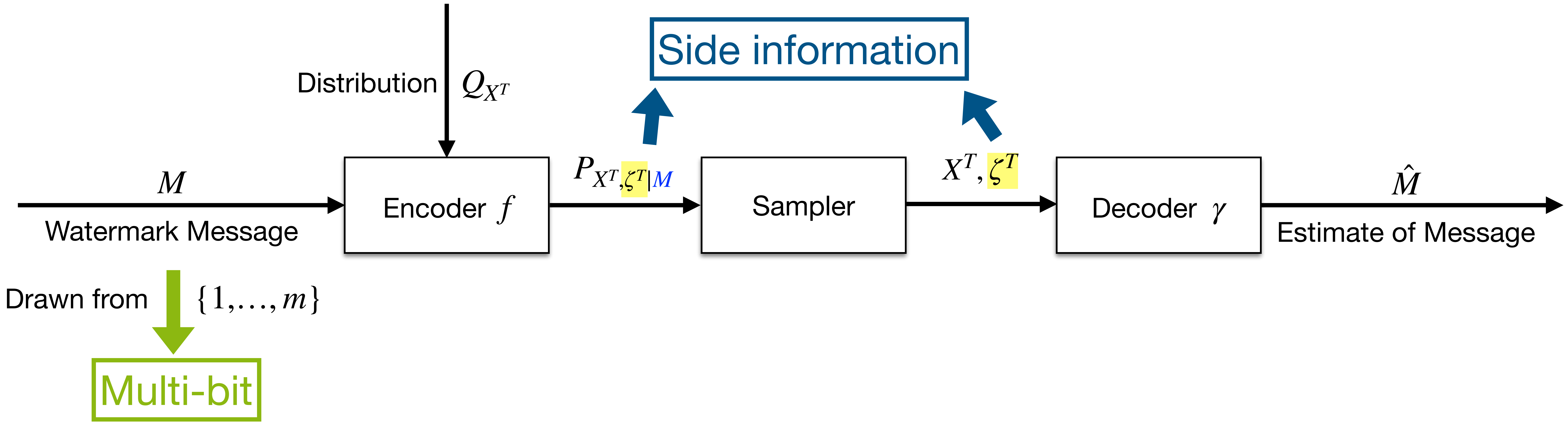
Distributional Information Embedding with Side Information——Multi-bit Watermarking



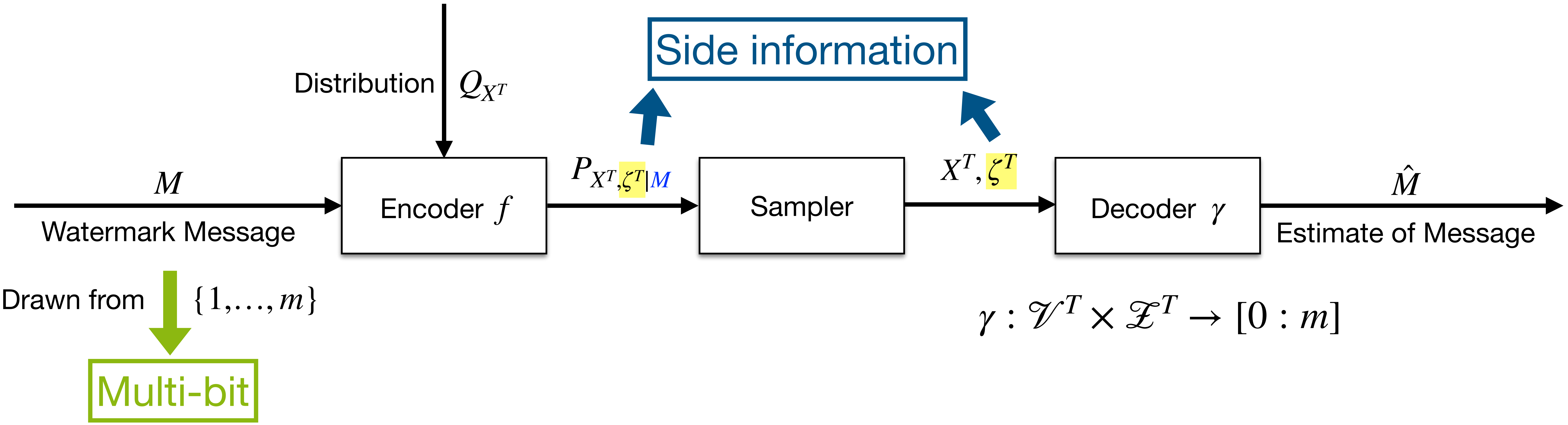
Distributional Information Embedding with Side Information——Multi-bit Watermarking



Distributional Information Embedding with Side Information——Multi-bit Watermarking



Distributional Information Embedding with Side Information——Multi-bit Watermarking



Distribution Q_{X^T}

Side information

M

Watermark Message

Encoder f

$P_{X^T,\zeta^T|M}$

Sampler

X^T, ζ^T

Decoder γ

$\hat{M}$

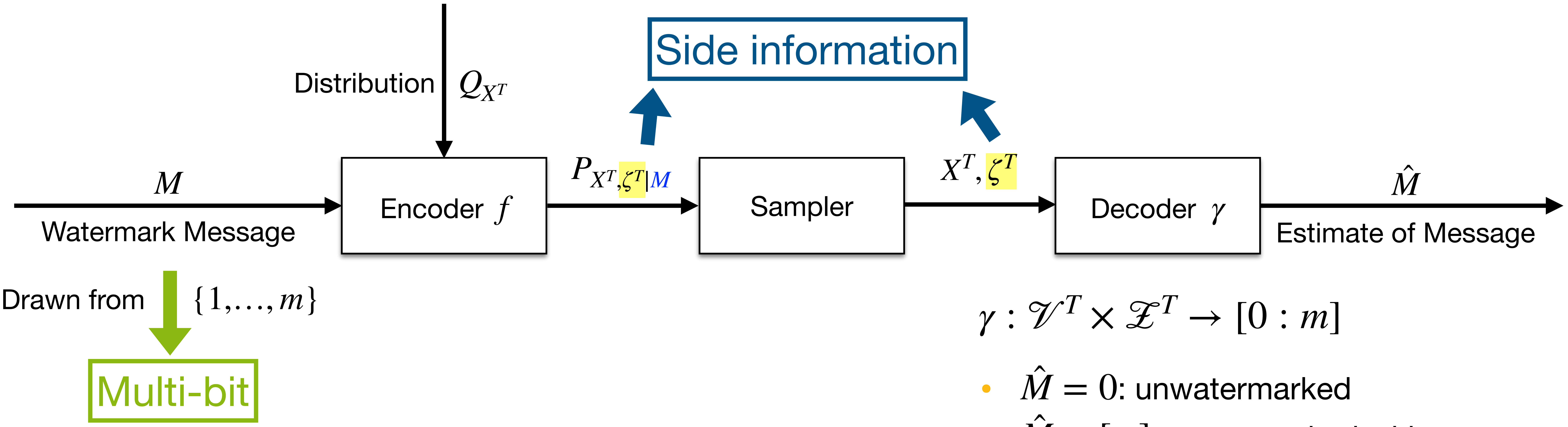
Estimate of Message

Drawn from $\{1,\ldots,m\}$

Multi-bit

$$\gamma : \mathscr{V}^T \times \mathscr{Z}^T \to [0:m]$$

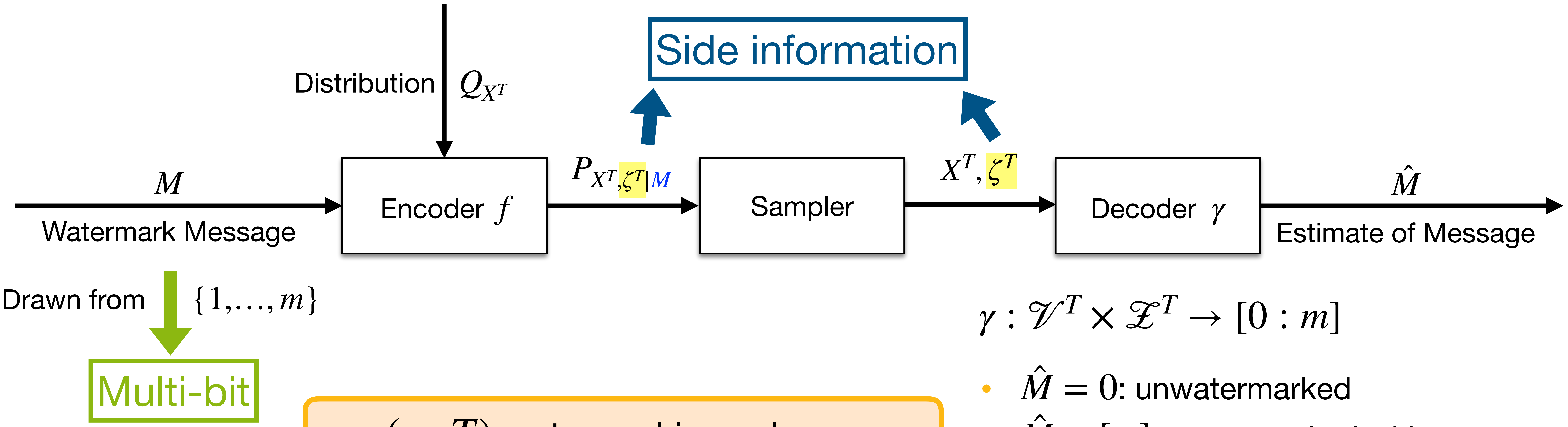
Distributional Information Embedding with Side Information——Multi-bit Watermarking



$$\gamma : \mathscr{V}^T \times \mathscr{Z}^T \to [0:m]$$

- $\hat{M} = 0$: unwatermarked
- $\hat{M} \in [m]$: watermarked with message $\hat{M}$

Distributional Information Embedding with Side Information——Multi-bit Watermarking



Side information

Distribution Q_{X^T}

M

Watermark Message

Encoder f

$P_{X^T,\zeta^T|M}$

Sampler

X^T, ζ^T

Decoder γ

$\hat{M}$

Estimate of Message

Drawn from $\{1,\ldots,m\}$

Multi-bit

$$\gamma : \mathscr{V}^T \times \mathscr{Z}^T \to [0:m]$$

- $\hat{M} = 0$: unwatermarked
- $\hat{M} \in [m]$: watermarked with message $\hat{M}$

(m, T) watermarking scheme

with **information rate** $R = \dfrac{\log m}{T}$

Secrecy of Embedded Message

Assumption 1

The encoder f must ensure that both X^T and ζ^T are statistically independent of message M.

Secrecy of Embedded Message

Assumption 1

The encoder f must ensure that both X^T and ζ^T are statistically independent of message M.

- Message M cannot be inferred simply from X^T or ζ^T

Secrecy of Embedded Message

Assumption 1

The encoder f must ensure that both X^T and ζ^T are statistically independent of message M.

- Message M cannot be inferred simply from X^T or ζ^T
- Must exploit the joint structure

$$\mathtt{I}(M; X^T, \zeta^T) = \mathtt{I}(M; X^T \,|\, \zeta^T) = \mathtt{I}(M; \zeta^T \,|\, X^T)$$

Multi-bit Watermarked Text Quality

Multi-bit Watermarked Text Quality

watermarked text distribution
with embedded message M

$$P_{X^T|M} = P_{X^T}$$

Multi-bit Watermarked Text Quality

watermarked text distribution
with embedded message M
$P_{X^T|M} = P_{X^T}$

vs

original text distribution
Q_{X^T}

Multi-bit Watermarked Text Quality

watermarked text distribution
with embedded message M

$$P_{X^T|M} = P_{X^T}$$

vs

original text distribution

$$Q_{X^T}$$

Good text quality

Multi-bit Watermarked Text Quality

watermarked text distribution
with embedded message M

$$P_{X^T|M} = P_{X^T}$$

vs

original text distribution

$$Q_{X^T}$$

Good text quality ⇔ $\mathsf{D}(P_{X^T}, Q_{X^T}) \leq d$

Multi-bit Watermarked Text Quality

watermarked text distribution
with embedded message M

$$P_{X^T|M} = P_{X^T}$$

vs

original text distribution

$$Q_{X^T}$$

Good text quality ⇔ $\mathrm{D}(P_{X^T}, Q_{X^T}) \leq d$

↓

(Distortion Level)

Multi-bit Watermarked Text Quality

watermarked text distribution
with embedded message M
$P_{X^T|M} = P_{X^T}$

vs

original text distribution
Q_{X^T}

Good text quality ⟺ $\mathrm{D}(P_{X^T}, Q_{X^T}) \leq d$ (D can be any distortion metric)

↓
(Distortion Level)

LLM Multi-bit Watermark Detection

Watermark Detection $\Longrightarrow$ $(m+1)$-ary Hypothesis Testing:

$\mathrm{H}_0 : X^T$ is human written, i.e., $(X^T, \zeta^T) \sim \mathbb{P}_j \triangleq Q_{X^T} \otimes P_{\zeta^T}$

$\mathrm{H}_j, \forall j \in [m] : X^T$ is LLM generated with embedded message j, i.e., $(X^T, \zeta^T) \sim \mathbb{P}_j \triangleq P_{X^T, \zeta^T | M=j}$

LLM Multi-bit Watermark Detection

Watermark Detection $\Longrightarrow$ $(m+1)$-ary Hypothesis Testing:

Human/unwatermarked LLM

H_0 : X^T is human written, i.e., $(X^T, \zeta^T) \sim \mathbb{P}_j \triangleq \boxed{Q_{X^T}} \otimes P_{\zeta^T}$

$\mathrm{H}_j, \forall j \in [m]$: X^T is LLM generated with embedded message j, i.e., $(X^T, \zeta^T) \sim \mathbb{P}_j \triangleq P_{X^T, \zeta^T | M=j}$

LLM Multi-bit Watermark Detection

Watermark Detection $\Longrightarrow$ $(m+1)$-ary Hypothesis Testing:

Human/unwatermarked LLM

$\mathrm{H}_0 : X^T$ is human written, i.e., $(X^T, \zeta^T) \sim \mathbb{P}_j \triangleq \boxed{Q_{X^T}} \otimes \boxed{P_{\zeta^T}}$

$\mathrm{H}_j, \forall j \in [m] : X^T$ is LLM generated with embedded message j,

i.e., $(X^T, \zeta^T) \sim \mathbb{P}_j \triangleq \boxed{P_{X^T, \zeta^T | M=j}}$

Watermarking scheme

LLM Multi-bit Watermark Detection

Watermark Detection $\Longrightarrow$ $(m+1)$-ary Hypothesis Testing:

Human/unwatermarked LLM

$\mathrm{H}_0 : X^T$ is human written, i.e., $(X^T, \zeta^T) \sim \mathbb{P}_j \triangleq \boxed{Q_{X^T}} \otimes \boxed{P_{\zeta^T}}$

$\mathrm{H}_j, \forall j \in [m] : X^T$ is LLM generated with embedded message j,

i.e., $(X^T, \zeta^T) \sim \mathbb{P}_j \triangleq \boxed{P_{X^T, \zeta^T | M=j}}$ → Watermarking scheme

Performance metric: **false-alarm and j-th error probability**

LLM Multi-bit Watermark Detection

Watermark Detection $\Longrightarrow$ $(m+1)$-ary Hypothesis Testing:

Human/unwatermarked LLM

$\mathrm{H}_0 : X^T$ is human written, i.e., $(X^T, \zeta^T) \sim \mathbb{P}_j \triangleq \boxed{Q_{X^T}} \otimes \boxed{P_{\zeta^T}}$

$\mathrm{H}_j, \forall j \in [m] : X^T$ is LLM generated with embedded message j,

i.e., $(X^T, \zeta^T) \sim \mathbb{P}_j \triangleq \boxed{P_{X^T,\zeta^T|M=j}}$ Watermarking scheme

Performance metric: **false-alarm and j-th error probability**

$$FA(\gamma, Q_{X^T}, P_{\zeta^T}) = \mathbb{P}_0(\gamma(X_1^T, \zeta_1^T) \neq 0)$$

LLM Multi-bit Watermark Detection

Watermark Detection $\Longrightarrow$ $(m+1)$-ary Hypothesis Testing:

Human/unwatermarked LLM

H_0 : X^T is human written, i.e., $(X^T, \zeta^T) \sim \mathbb{P}_j \triangleq \boxed{Q_{X^T}} \otimes \boxed{P_{\zeta^T}}$

$\mathrm{H}_j, \forall j \in [m]$: X^T is LLM generated with embedded message j,

i.e., $(X^T, \zeta^T) \sim \mathbb{P}_j \triangleq \boxed{P_{X^T, \zeta^T | M=j}}$ Watermarking scheme

<u>Performance metric:</u> false-alarm and j-th error probability

$$FA(\gamma, Q_{X^T}, P_{\zeta^T}) = \mathbb{P}_0(\gamma(X_1^T, \zeta_1^T) \neq 0)$$

$$MD_j(\gamma, P_{X^T, \zeta^T | M=j}) = \mathbb{P}_j(\gamma(X_1^T, \zeta_1^T) \neq j)$$

Multi-bit Watermarking Design Objective

Three-fold

1. Maximize information rate $R = \dfrac{\log m}{T}$
2. Ensure text quality $\mathrm{D}(P_{X^T}, Q_{X^T}) \leq d$
3. Minimize MD_j while worst-case false alarm $\sup_{Q_{X^T}} FA \leq \alpha, \quad \forall j \in [m]$

Maximum Information Rate

Maximum Information Rate

- Asymptotic analysis when $T \to \infty$ for **IID processes**

Maximum Information Rate

- Asymptotic analysis when $T \to \infty$ for **IID processes**
- Assume $X_t \overset{iid}{\sim} P_X$, $\zeta_t \overset{iid}{\sim} P_\zeta$, and under H_j, $(X_t, \zeta_t) \overset{iid}{\sim} P_{X,\zeta|M=j}$

Maximum Information Rate

- Asymptotic analysis when $T \to \infty$ for **IID processes**
- Assume $X_t \overset{iid}{\sim} P_X$, $\zeta_t \overset{iid}{\sim} P_\zeta$, and under H_j, $(X_t, \zeta_t) \overset{iid}{\sim} P_{X,\zeta|M=j}$
- Assume uniform prior of message M

Maximum Information Rate

- Asymptotic analysis when $T \to \infty$ for **IID processes**
- Assume $X_t \overset{iid}{\sim} P_X$, $\zeta_t \overset{iid}{\sim} P_\zeta$, and under H_j, $(X_t, \zeta_t) \overset{iid}{\sim} P_{X,\zeta|M=j}$
- Assume uniform prior of message M

Lemma 1 (Maximum Information Rate)

If the decoding error $\Pr(\hat{M} \neq M) = \dfrac{1}{m}\sum_{j=1}^{m} MD_j \to 0$ as $T \to \infty$,

then we have $R \leq \sup_{P_X : \mathrm{D}(P_X^T, Q_X^T) \leq d} \mathrm{H}(P_X)$.

Maximum Information Rate

- Asymptotic analysis when $T \to \infty$ for **IID processes**
- Assume $X_t \overset{iid}{\sim} P_X$, $\zeta_t \overset{iid}{\sim} P_\zeta$, and under H_j, $(X_t, \zeta_t) \overset{iid}{\sim} P_{X,\zeta|M=j}$
- Assume uniform prior of message M

Lemma 1 (Maximum Information Rate)

If the decoding error $\Pr(\hat{M} \neq M) = \dfrac{1}{m}\sum_{j=1}^{m} MD_j \to 0$ as $T \to \infty$,

then we have $R \leq \sup_{P_X : \mathrm{D}(P_X^T, Q_X^T) \leq d} \mathrm{H}(P_X)$.

(X^T, ζ^T) stationary ergodic processes —> entropy rate

Asymptotically Optimal Multi-bit Watermarking

Asymptotically Optimal Multi-bit Watermarking

- **Goal:** vanishing detection error & maximum information rate

Asymptotically Optimal Multi-bit Watermarking

- **Goal:** vanishing detection error & maximum information rate
- Inspired by the upper bound of j-th error exponent

$$E_j^* = \max_{P_X:\mathrm{D}(P_X^T,Q_X^T)\leq d} \min_{i\in[0:m]\backslash j} \mathrm{D}_{\mathrm{KL}}(P_{X,\zeta|M=i}||P_{X,\zeta|M=j})$$

Asymptotically Optimal Multi-bit Watermarking

- **Goal:** vanishing detection error & maximum information rate
- Inspired by the upper bound of j-th error exponent

$$E_j^* = \max_{P_X : \mathrm{D}(P_X^T, Q_X^T) \leq d} \ \min_{i \in [0:m] \backslash j} \mathrm{D}_{\mathrm{KL}}(P_{X,\zeta|M=i} || P_{X,\zeta|M=j})$$

Design idea: make them concentrated at different locations

Asymptotically Optimal Multi-bit Watermarking

- **Goal:** vanishing detection error & maximum information rate

If $\frac{1}{T}(\log m - \log \alpha) \leq \sup_{\mathrm{D}(P_X^T, Q_X^T) \leq d} \mathrm{H}(P_X)$, we have

Asymptotically Optimal Multi-bit Watermarking

- **Goal:** vanishing detection error & maximum information rate

If $\frac{1}{T}(\log m - \log \alpha) \leq \sup_{\mathrm{D}(P_X^T, Q_X^T) \leq d} \mathrm{H}(P_X)$, we have

(Example: m=3)

Detector γ^*

Encoder output $P^{*T}_{X,\zeta|M}$

Asymptotically Optimal Multi-bit Watermarking

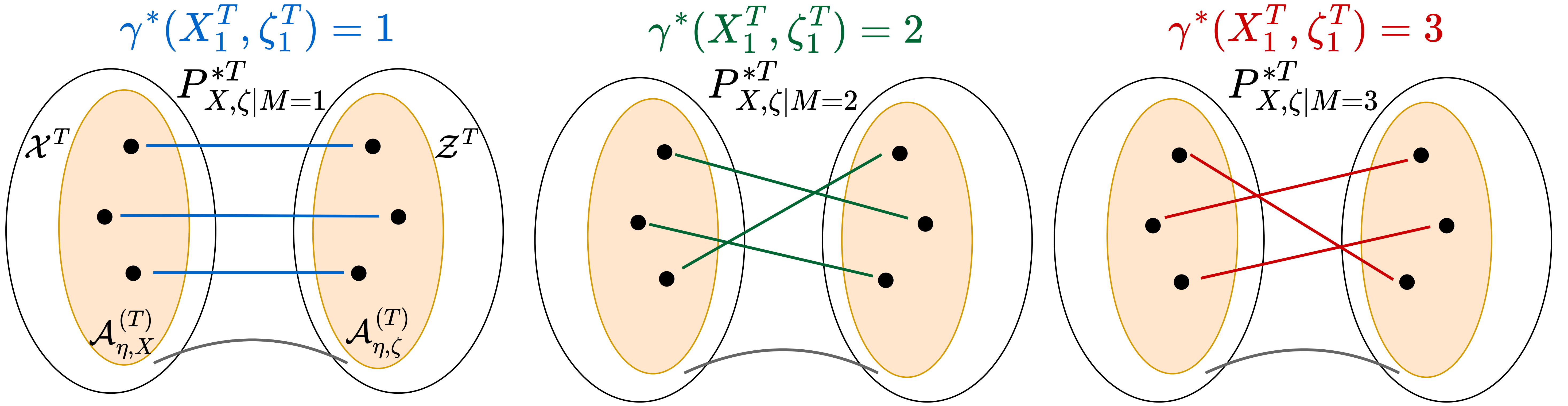
- **Goal:** vanishing detection error & maximum information rate

If $\frac{1}{T}(\log m - \log \alpha) \leq \sup_{\mathsf{D}(P_X^T, Q_X^T) \leq d} \mathsf{H}(P_X)$, we have

(Example: m=3)

Detector γ^*

Encoder output $P^{*T}_{X,\zeta|M}$


$\gamma^*(X_1^T, \zeta_1^T) = 1$
$P^{*T}_{X,\zeta|M=1}$
$\mathcal{X}^T$
$\mathcal{Z}^T$
$\mathcal{A}^{(T)}_{\eta,X}$
$\mathcal{A}^{(T)}_{\eta,\zeta}$
$\gamma^*(X_1^T, \zeta_1^T) = 2$
$P^{*T}_{X,\zeta|M=2}$
$\gamma^*(X_1^T, \zeta_1^T) = 3$
$P^{*T}_{X,\zeta|M=3}$

Asymptotically Optimal Multi-bit Watermarking

- **Goal:** vanishing detection error & maximum information rate

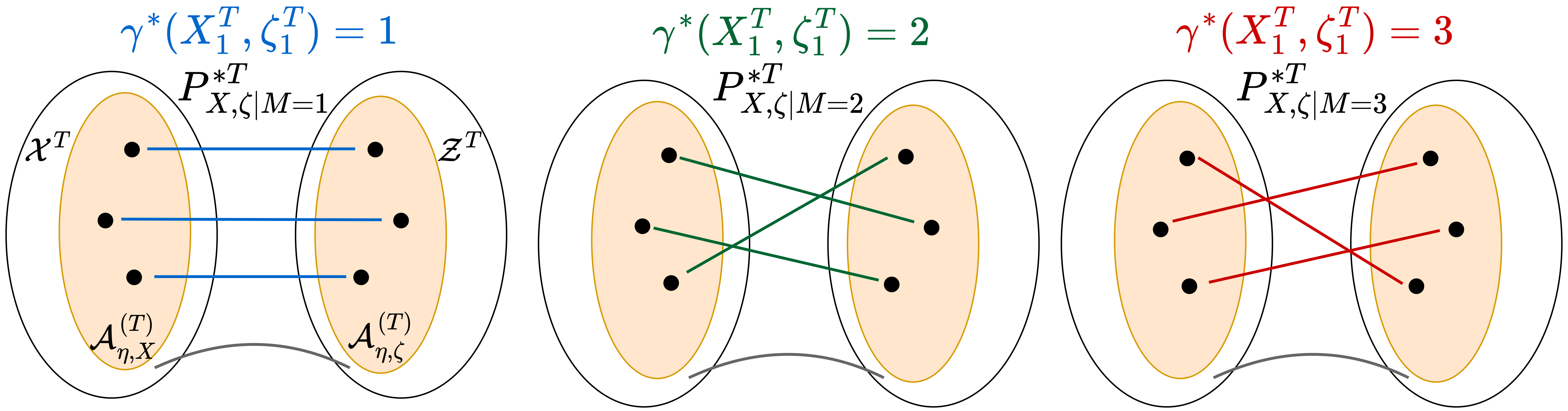
If $\frac{1}{T}(\log m - \log \alpha) \le \sup\limits_{\mathsf{D}(P_X^T, Q_X^T) \le d} \mathsf{H}(P_X)$, we have

(Example: m=3)

Detector γ^*

Encoder output $P^{*T}_{X,\zeta|M}$

$$P_X^* = \arg \max_{P_X : \mathsf{D}(P_X^T, Q_X^T) \le d} \mathsf{H}(P_X)$$


$\gamma^*(X_1^T, \zeta_1^T) = 1$
$P^{*T}_{X,\zeta|M=1}$
$\mathcal{X}^T$
$\mathcal{Z}^T$
$\mathcal{A}^{(T)}_{\eta,X}$
$\mathcal{A}^{(T)}_{\eta,\zeta}$
$\gamma^*(X_1^T, \zeta_1^T) = 2$
$P^{*T}_{X,\zeta|M=2}$
$\gamma^*(X_1^T, \zeta_1^T) = 3$
$P^{*T}_{X,\zeta|M=3}$

Asymptotically Optimal Multi-bit Watermarking

- **Goal:** vanishing detection error & maximum information rate

If $\frac{1}{T}(\log m - \log \alpha) \leq \sup_{\mathrm{D}(P_X^T, Q_X^T) \leq d} \mathrm{H}(P_X)$, we have

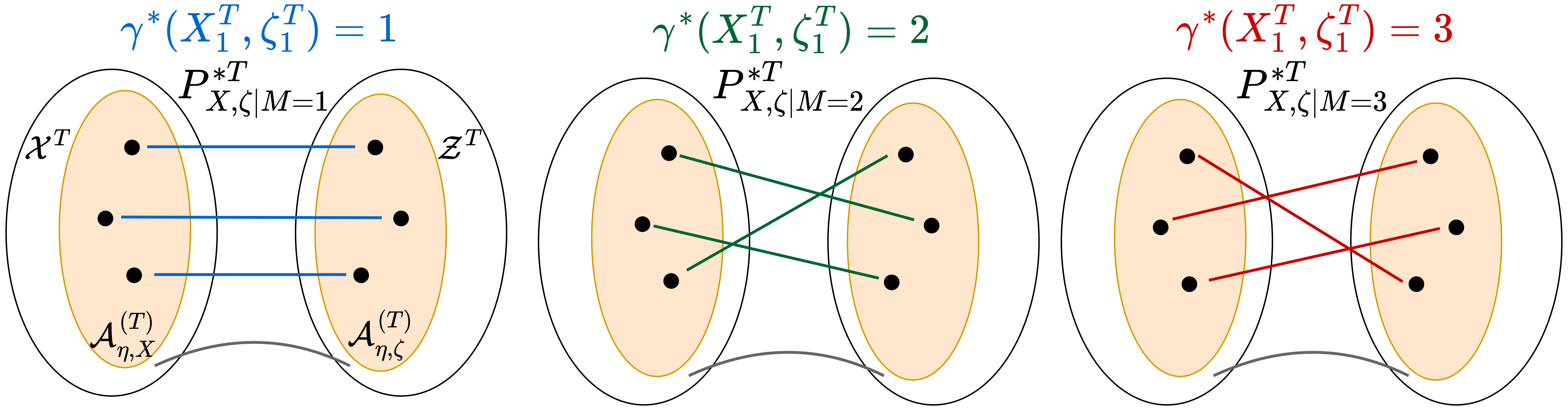
(Example: m=3)

Detector γ^*

Encoder output $P^{*T}_{X,\zeta|M}$

$P_X^* = \arg \max_{P_X : \mathrm{D}(P_X^T, Q_X^T) \leq d} \mathrm{H}(P_X)$

P_ζ^*: $\mathrm{H}(P_\zeta^*) = \mathrm{H}(P_X^*)$



$\gamma^*(X_1^T, \zeta_1^T) = 1$
$P^{*T}_{X,\zeta|M=1}$
$\mathcal{X}^T$
$\mathcal{Z}^T$
$\mathcal{A}^{(T)}_{\eta,X}$
$\mathcal{A}^{(T)}_{\eta,\zeta}$
$\gamma^*(X_1^T, \zeta_1^T) = 2$
$P^{*T}_{X,\zeta|M=2}$
$\gamma^*(X_1^T, \zeta_1^T) = 3$
$P^{*T}_{X,\zeta|M=3}$

Asymptotically Optimal Multi-bit Watermarking

- **Goal:** vanishing detection error & maximum information rate

If $\frac{1}{T}(\log m - \log \alpha) \le \sup_{\mathsf{D}(P_X^T, Q_X^T)\le d} \mathsf{H}(P_X)$, we have

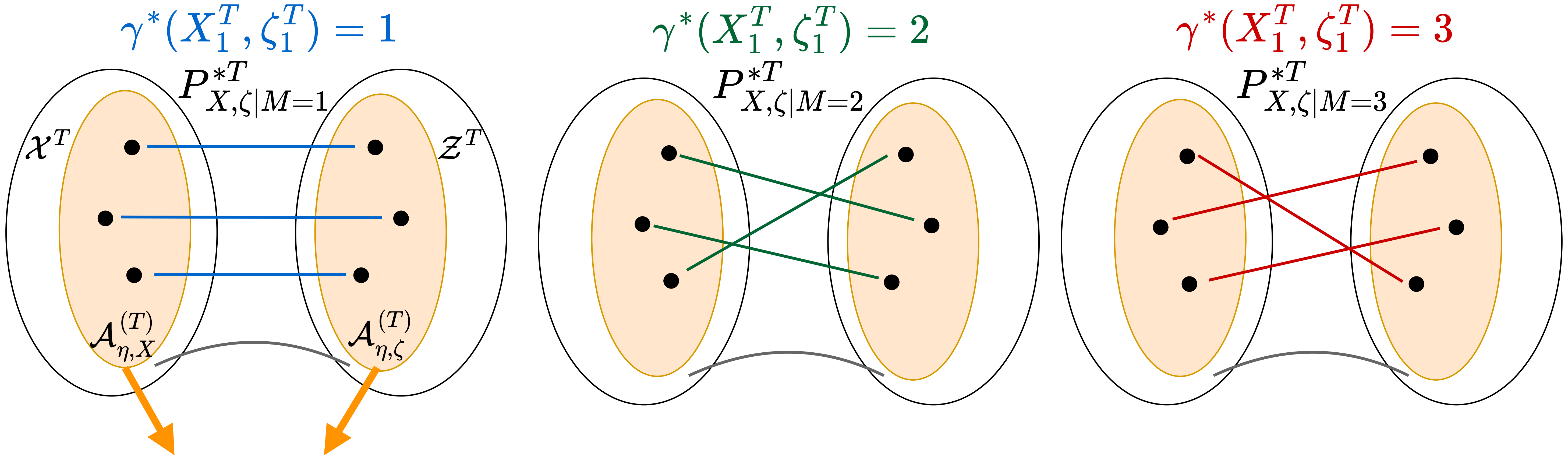
(Example: m=3)

Detector γ^*

Encoder output $P^{*T}_{X,\zeta|M}$

$P_X^* = \arg\max_{P_X:\mathsf{D}(P_X^T, Q_X^T)\le d} \mathsf{H}(P_X)$

P_ζ^*: $\mathsf{H}(P_\zeta^*) = \mathsf{H}(P_X^*)$



$\gamma^*(X_1^T, \zeta_1^T) = 1$
$P^{*T}_{X,\zeta|M=1}$
$\mathcal{X}^T$
$\mathcal{Z}^T$
$\mathcal{A}^{(T)}_{\eta,X}$
$\mathcal{A}^{(T)}_{\eta,\zeta}$
$\gamma^*(X_1^T, \zeta_1^T) = 2$
$P^{*T}_{X,\zeta|M=2}$
$\gamma^*(X_1^T, \zeta_1^T) = 3$
$P^{*T}_{X,\zeta|M=3}$

Typical set

Asymptotically Optimal Multi-bit Watermarking

- **Goal:** vanishing detection error & maximum information rate

If $\dfrac{1}{T}(\log m - \log \alpha) \leq \sup\limits_{\mathsf{D}(P_X^T, Q_X^T) \leq d} \mathsf{H}(P_X)$, we have

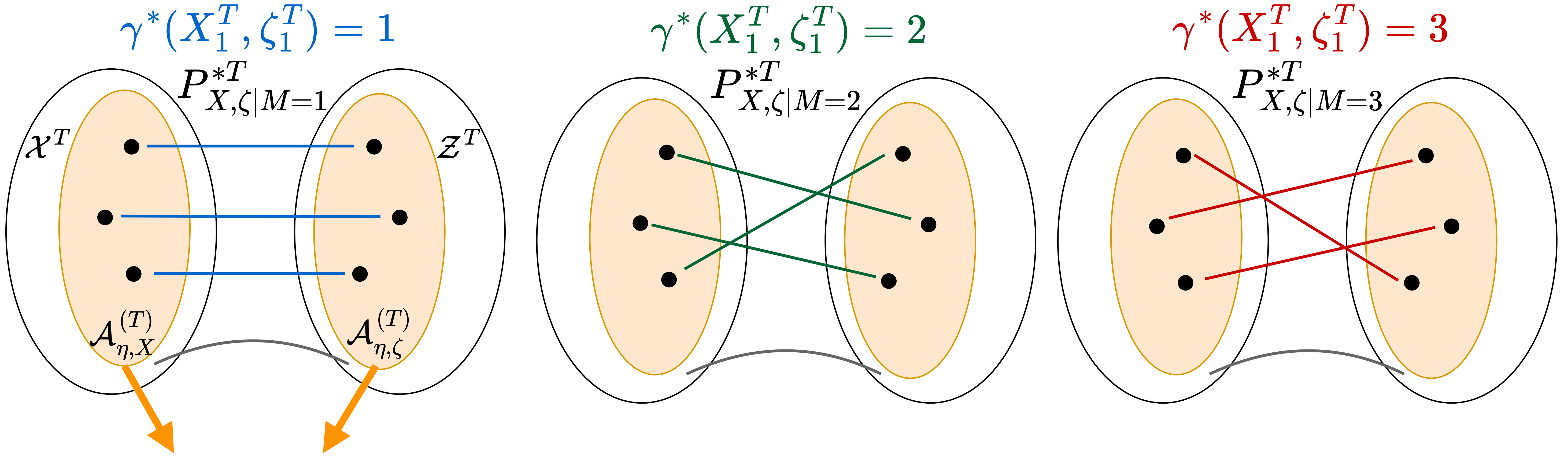
(Example: m=3)

Detector γ^*

Encoder output $P^{*T}_{X,\zeta|M}$

$P_X^* = \arg \max\limits_{P_X : \mathsf{D}(P_X^T, Q_X^T) \leq d} \mathsf{H}(P_X)$

P_ζ^*: $\mathsf{H}(P_\zeta^*) = \mathsf{H}(P_X^*)$



$\gamma^*(X_1^T, \zeta_1^T) = 1$
$P^{*T}_{X,\zeta|M=1}$
$\mathcal{X}^T$
$\mathcal{Z}^T$
$\mathcal{A}_{\eta,X}^{(T)}$
$\mathcal{A}_{\eta,\zeta}^{(T)}$
$\gamma^*(X_1^T, \zeta_1^T) = 2$
$P^{*T}_{X,\zeta|M=2}$
$\gamma^*(X_1^T, \zeta_1^T) = 3$
$P^{*T}_{X,\zeta|M=3}$

Typical set

$\gamma^* = 0$ if (x^T, ζ^T) are atypical

Asymptotically Optimal Multi-bit Watermarking

- **Goal:** vanishing detection error & maximum information rate

If $\frac{1}{T}(\log m - \log \alpha) \leq \sup_{\mathrm{D}(P_X^T, Q_X^T) \leq d} \mathrm{H}(P_X)$, we have

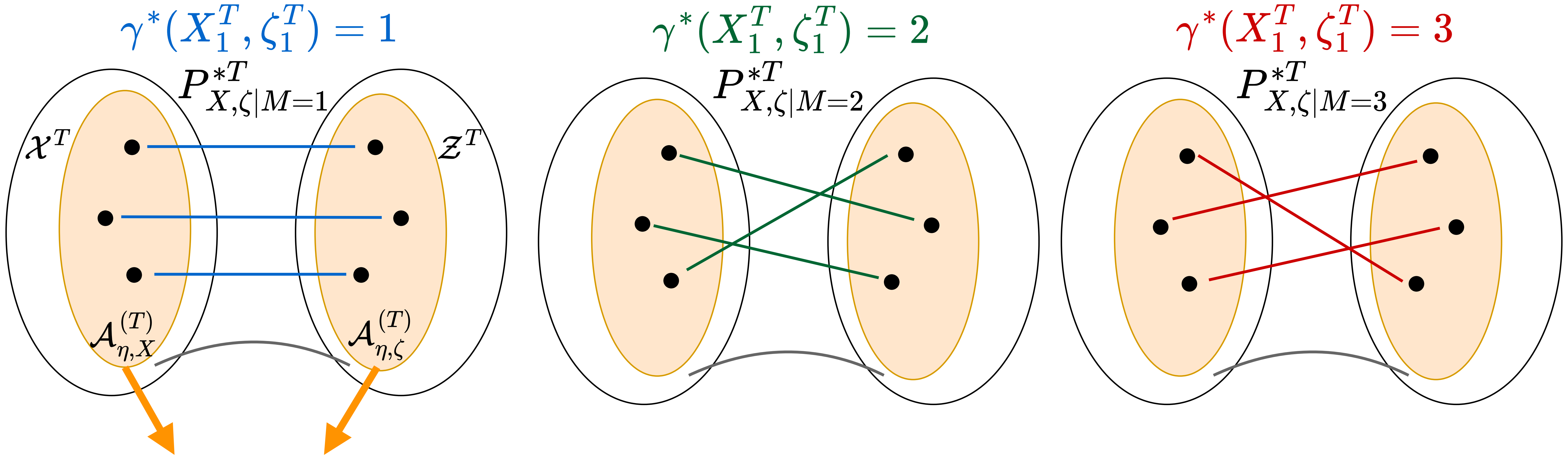
(Example: m=3)

Detector γ^*

Encoder output $P^{*T}_{X,\zeta|M}$

$P_X^* = \arg \max_{P_X : \mathrm{D}(P_X^T, Q_X^T) \leq d} \mathrm{H}(P_X)$

P_ζ^*: $\mathrm{H}(P_\zeta^*) = \mathrm{H}(P_X^*)$



$\gamma^*(X_1^T, \zeta_1^T) = 1$
$P^{*T}_{X,\zeta|M=1}$
$\mathcal{X}^T$
$\mathcal{Z}^T$
$\mathcal{A}^{(T)}_{\eta,X}$
$\mathcal{A}^{(T)}_{\eta,\zeta}$
$\gamma^*(X_1^T, \zeta_1^T) = 2$
$P^{*T}_{X,\zeta|M=2}$
$\gamma^*(X_1^T, \zeta_1^T) = 3$
$P^{*T}_{X,\zeta|M=3}$

Typical set

$\gamma^* = 0$ if (x^T, ζ^T) are atypical

This ensures: $\forall j \in [m], MD_j \to 0,\ FA \to 0$, **and max** $R \to \sup_{\mathrm{D}(P_X^T, Q_X^T) \leq d} \mathrm{H}(P_X)$

Finite-Length Analysis

Optimization problem:

$$\min_{\gamma,\, P_{X^T,\zeta^T|M=j}} \quad MD_j(\gamma,\, P_{X^T,\zeta^T|M=j})$$

$$\text{s.t.} \quad \sup_{P_{X^T,\zeta^T|M=i}} MD_i(\gamma,\, P_{X^T,\zeta^T|M=i}) \leq \alpha, \quad \forall i \neq j$$

$$\sup_{Q_{X^T}} FA(\gamma, Q_{X^T}, P_{\zeta^T}) \leq \alpha$$

$$\mathrm{D}(P_{X^T}, Q_{X^T}) \leq \epsilon$$

Finite-Length Analysis

Optimization problem:

$$\min_{\gamma,\, P_{X^T,\zeta^T|M=j}} \quad MD_j(\gamma,\, P_{X^T,\zeta^T|M=j})$$

s.t. $$\sup_{P_{X^T,\zeta^T|M=i}} MD_i(\gamma,\, P_{X^T,\zeta^T|M=i}) \leq \alpha, \quad \forall i \neq j$$

$$\sup_{Q_{X^T}} FA(\gamma, Q_{X^T}, P_{\zeta^T}) \leq \alpha$$

$$\mathrm{D}(P_{X^T}, Q_{X^T}) \leq \epsilon$$

✦ **Lower bound on MD_j:**

$$MD_j \geq m\beta^*(\alpha, T),$$

where

$$\beta^*(\alpha, T) = \sum_{x^T} \left(P^*_{X^T}(x^T) - \alpha\right)_+$$

Finite-Length Analysis

Optimization problem:

$$\min_{\gamma,\, P_{X^T,\zeta^T|M=j}} \quad MD_j(\gamma,\, P_{X^T,\zeta^T|M=j})$$

s.t. $\sup_{P_{X^T,\zeta^T|M=i}} MD_i(\gamma,\, P_{X^T,\zeta^T|M=i}) \leq \alpha, \quad \forall i \neq j$

$\sup_{Q_{X^T}} FA(\gamma, Q_{X^T}, P_{\zeta^T}) \leq \alpha$

$\mathrm{D}(P_{X^T}, Q_{X^T}) \leq \epsilon$

There are m problems in total.

✦ **Lower bound on MD_j:**

$$MD_j \geq m\beta^*(\alpha, T),$$

where

$$\beta^*(\alpha, T) = \sum\nolimits_{x^T} \left(P^*_{X^T}(x^T) - \alpha\right)_+$$

Finite-Length Analysis

Optimization problem:

$$\min_{\gamma,\, P_{X^T,\zeta^T|M=j}} \quad MD_j(\gamma,\, P_{X^T,\zeta^T|M=j})$$

$$\text{s.t.} \quad \sup_{P_{X^T,\zeta^T|M=i}} MD_i(\gamma,\, P_{X^T,\zeta^T|M=i}) \le \alpha, \quad \forall i \neq j$$

$$\sup_{Q_{X^T}} FA(\gamma, Q_{X^T}, P_{\zeta^T}) \le \alpha$$

$$\mathrm{D}(P_{X^T}, Q_{X^T}) \le \epsilon$$

There are m problems in total.

✦ **Lower bound on MD_j:**

$$MD_j \ge m\beta^*(\alpha, T),$$

where

$$\beta^*(\alpha, T) = \sum\nolimits_{x^T} \left(P^*_{X^T}(x^T) - \alpha\right)_+$$

⬇

$$m \le 1/\beta^*(\alpha, T)$$

Finite-Length Analysis

Optimization problem:

$$\min_{\gamma,\, P_{X^T,\zeta^T|M=j}} \quad MD_j(\gamma,\, P_{X^T,\zeta^T|M=j})$$

$$\text{s.t.} \quad \sup_{P_{X^T,\zeta^T|M=i}} MD_i(\gamma,\, P_{X^T,\zeta^T|M=i}) \le \alpha, \quad \forall i \ne j$$

$$\sup_{Q_{X^T}} FA(\gamma, Q_{X^T}, P_{\zeta^T}) \le \alpha$$

$$\mathrm{D}(P_{X^T}, Q_{X^T}) \le \epsilon$$

There are m problems in total.

✦ **Lower bound on MD_j:**

$$MD_j \ge m\beta^*(\alpha, T),$$

where

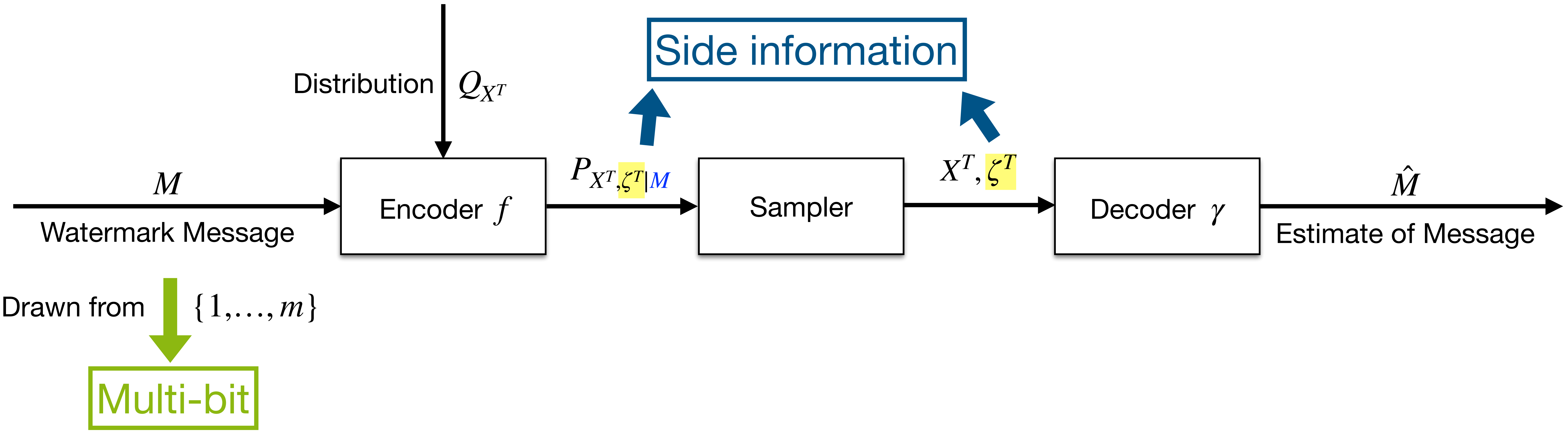
$$\beta^*(\alpha, T) = \sum_{x^T} \left(P^*_{X^T}(x^T) - \alpha\right)_+$$

⬇

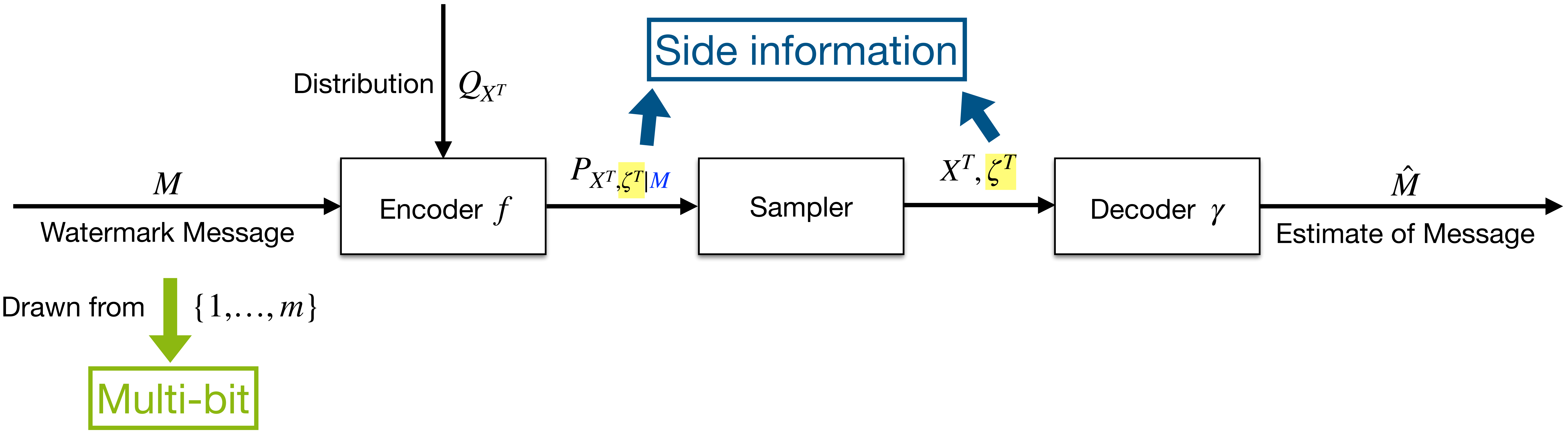
$$m \le 1/\beta^*(\alpha, T)$$

Achievability: future work

Summary


Side information
Distribution Q_{X^T}
M
Watermark Message
Encoder f
$P_{X^T,\zeta^T|M}$
Sampler
X^T, ζ^T
Decoder γ
$\hat{M}$
Estimate of Message
Drawn from $\{1, \ldots, m\}$
Multi-bit

Summary



Thank you! ☺