# Online Conformal Compression for Zero-Delay Communication with Distortion Guarantees

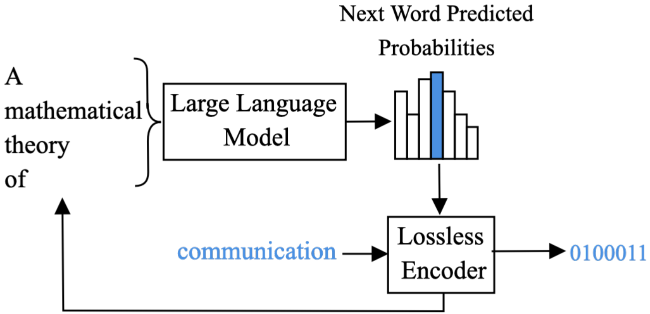
Unnikrishnan Kunnath Ganesan*, Giuseppe Durisi*, Matteo Zecchin†, Petar Popovski‡ and Osvaldo Simeone†

*Chalmers University of Technology, †Kings College London, ‡Aalborg University

Thursday 26th June, 2025

# Motivation

- Thanks to advancements in **deep sequence modeling**, a range of powerful predictive models are widely available.
- Deep sequence models can be combined with compression algorithms to realize powerful prediction-powered sequence compressors [1,2].

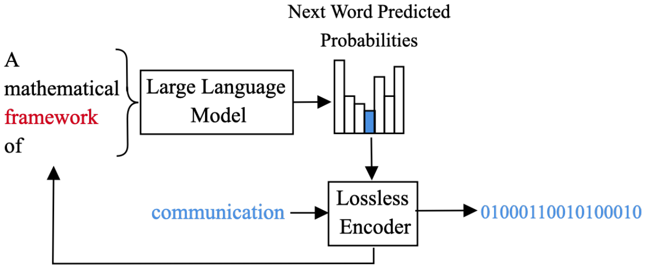


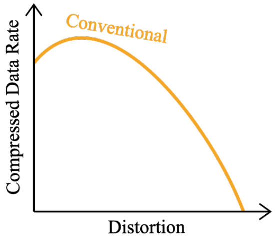

[1]Schmidhuber and Heil, “Sequential neural text compression”,IEEE Transactions on Neural Networks, 1996

[2]Mahoney, “Fast Text Compression with Neural Networks.”, Thirteenth International Florida Artificial Intelligence Research Society Conference, 2000

# Motivation

- Most compressors based on deep sequence models are **lossless**[3,4], using e.g., arithmetic codes.
- Distortion introduces errors in the predictor's input and can degrade the compression rate.



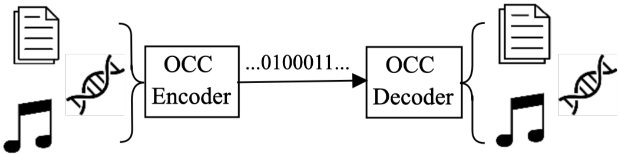



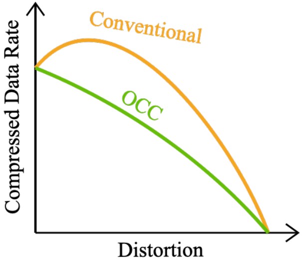

[3]Goyal et al., "Deepzip: Lossless data compression using recurrent neural networks", Data Compression Conference (DCC) 2019

[4]Valmeekam et al., "LLMzip: Lossless text compression using large language models", arXiv preprint, 2023

# Our contribution

- We introduce **online conformal compression** (OCC), a zero-delay, prediction-powered lossy compression scheme.
- OCC applies **online conformal prediction**[5], providing **deterministic distortion guarantees** for any input sequence, regardless of predictor quality.

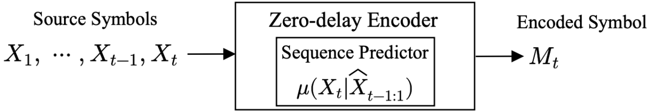


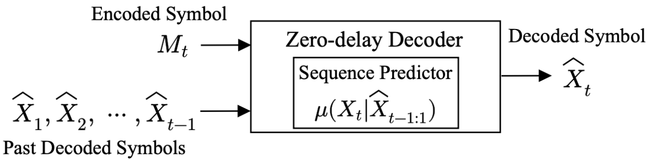

[5] Gibbs and Candes, "Adaptive conformal inference under distribution shift", NeurIPS, 2021

# Problem Formulation

- **Zero-delay encoder:** $X_1, \ldots, X_{t-1}, X_t \xrightarrow{\mu(X_t|\hat{X}_{t-1:1})} M_t$ ($b_t$-bits)




- **Zero-delay decoder:** $\hat{X}_1, \ldots, \hat{X}_{t-1}, M_t \xrightarrow{\mu(X_t|\hat{X}_{t-1:1})} \hat{X}_t$



# Problem Formulation

- Goal: Design a sequence of encoding/decoding functions that minimizes the **transmission rate**

$$B_T = \frac{1}{T}\sum_{t=1}^{T} b_t,$$

while guaranteeing for any $\alpha \in [0,1]$ the **deterministic distortion requirement**

$$\underbrace{\frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\{\hat{X}_t \neq X_t\}}_{\text{distortion}} \leq \underbrace{\alpha}_{\text{target}} + \frac{C}{T}$$

for some constant $C > 0$.

# Prediction Set-Based Encoding

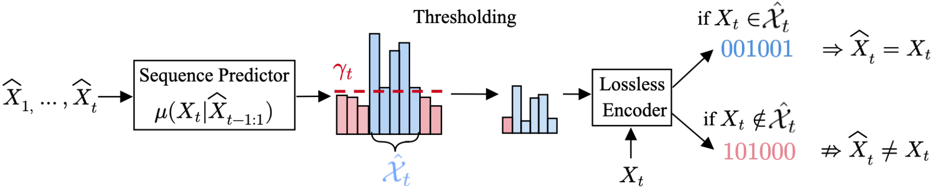
- Obtain the **high-probability prediction set**

$$\hat{\mathcal{X}}_t = \{X \in \mathcal{X} : \mu_t(X|\hat{X}_{t-1:1}) \geq \gamma_t\}.$$

- Use lossless compression for the truncated distribution

$$\bar{\mu}_t(x) = \frac{\mu_t(x|\hat{X}_{t-1:1})}{\sum_{x' \in \hat{\mathcal{X}}_t} \mu_t(x'|\hat{X}_{t-1:1})}, \quad \forall x \in \hat{\mathcal{X}}_t.$$

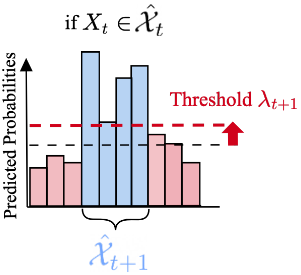
- An outage symbol is sent if $X_t \notin \hat{\mathcal{X}}_t$.

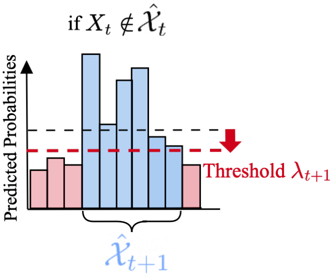


# Online Conformal Prediction

- Online conformal prediction[6] updates the threshold based on past errors as

$$\gamma_{t+1} = \gamma_t - \eta(\underbrace{\mathbb{1}\{X_t \notin \hat{\mathcal{X}}_t\}}_{\text{instantaneous miscoverage error}} - \alpha),$$




[6] Gibbs and Candes, "Adaptive conformal inference under distribution shift", NeurIPS, 2021

# Online Conformal Prediction

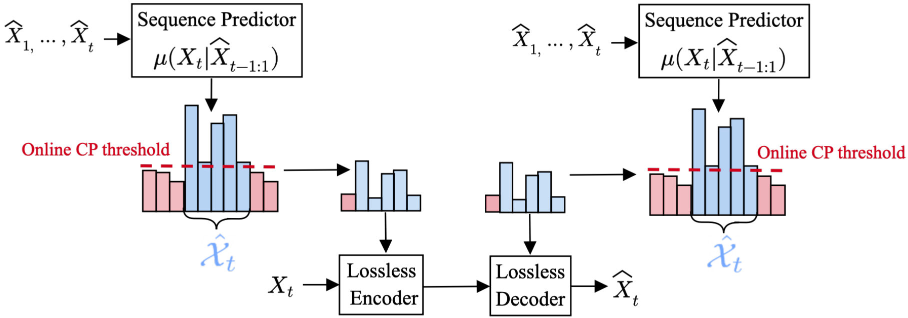
**Lemma**

*For any sequence $X_1, \dots, X_T$, the prediction sets $\hat{\mathcal{X}}_1, \dots, \hat{\mathcal{X}}_T$ satisfy the coverage guarantee*

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{1}\{X_t \notin \hat{\mathcal{X}}_t\} \le \alpha + \frac{1+\eta}{\eta T}.$$

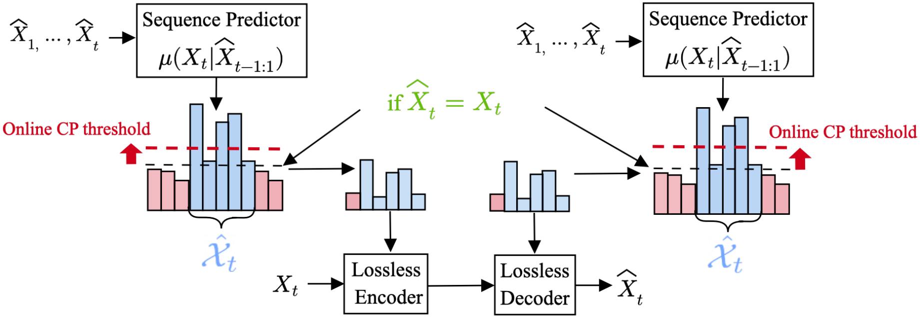
# Online Conformal Compression

- In **online conformal compression** (OCC), the encoder and decoder apply a prediction set-based encoding with threshold updated based on a variant of online conformal prediction.
- The encoded and the decoder use identical learned model for prediction which is pre-shared.



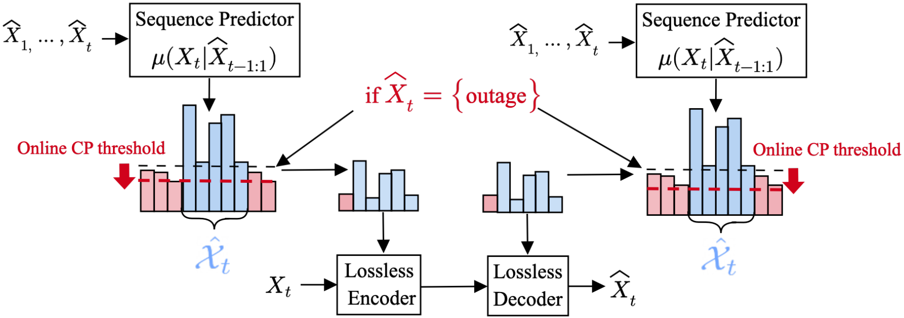
# Online Conformal Compression

**1)** If $X_t \in \mathcal{X}_t$, we have $\hat{X}_t = X_t$ by construction and the threshold $\gamma_t$ is increased as $\gamma_{t+1} = \gamma_t + \eta\alpha$



# Online Conformal Compression

**2)** If $X_t \notin \mathcal{X}_t$, we may have $\hat{X}_t \neq X_t$, and the threshold $\gamma_t$ is decreased as
$\gamma_{t+1} = \gamma_t - \eta(1 - \alpha)$



# Distortion Guarantees

## Theorem

*For any distortion level* $\alpha \in [0,1]$*, any sequence of predictors* $\mu_t(\cdot)$ *and any data sequence* $X_1, \ldots, X_T$*, the distortion of online conformal compression is deterministically bounded as*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\{\hat{X}_t \neq X_t\} \leq \alpha + \frac{1+\eta}{\eta T}.$$

## Proof.

Online conformal prediction's bound on the outage events

$$\Downarrow$$

Online conformal compression's bound on the distortion

$\square$

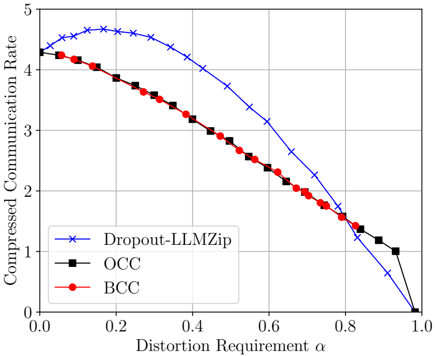
# Compression Performance

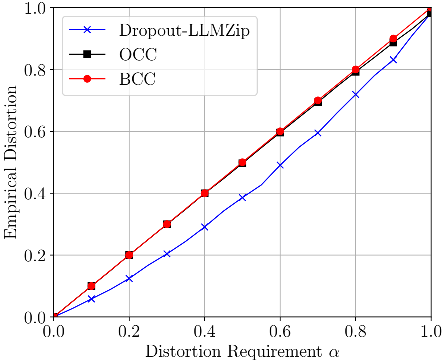
- Prediction model: nanoGPT model
- Data set: Shakespeare's works and Taylor Swift's songs.
- Baselines:
  - Dropout-LLMZip[7], applies lossless compression (i.e., no thresholding) and drops symbols with probability $\alpha$.
  - Block Conformal Compression (BCC), applies prediction set encoding with the **optimal fixed threshold** chosen in hindsight based on the entire sequence.

[7] Valmeekam et al., "LLMzip: Lossless text compression using large language models", arXiv preprint, 2023

# Stationary Data

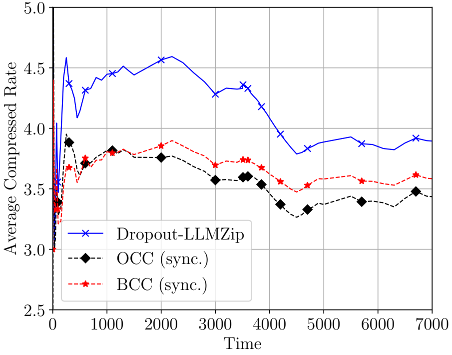
- OCC matches the performance of BCC and outperforms Dropout-LLMZip for most distortion levels.

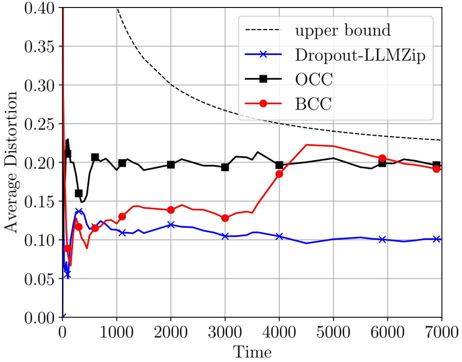


# Non-stationary Data

- Up to time $T = 3500$ data comes from Shakespeare's works and after from Taylor's swift songs.
- OCC adapts to the different data distribution and delivers an homogeneous distortion level over time.



# Conclusion

- We introduce online conformal compression (OCC), a zero-delay prediction-powered sequence compressor with deterministic distortion guarantees.
- OCC combines online conformal prediction with a prediction set-based compression to achieve performance comparable to that an offline encoder optimized in hindsight.