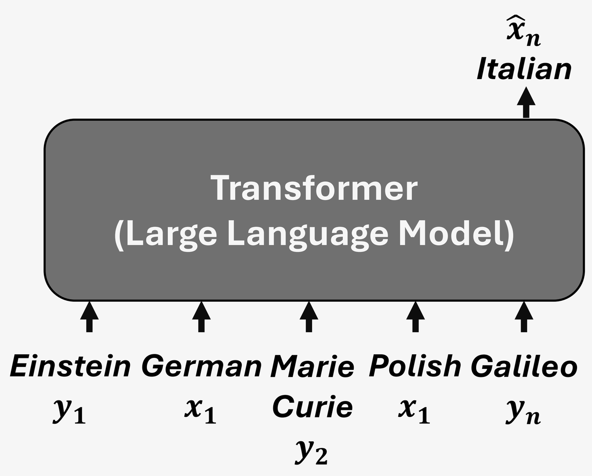
Transformers are Provably Efficient In-context Estimators for Wireless Communications

V. T. Kunde V. Rajagopalan C. S. K. Valmeekam J.-F. Chamberland K. Narayanan D. Kalathil S. Shakkottai

Department of Electrical and Computer Engineering, Texas A&M University

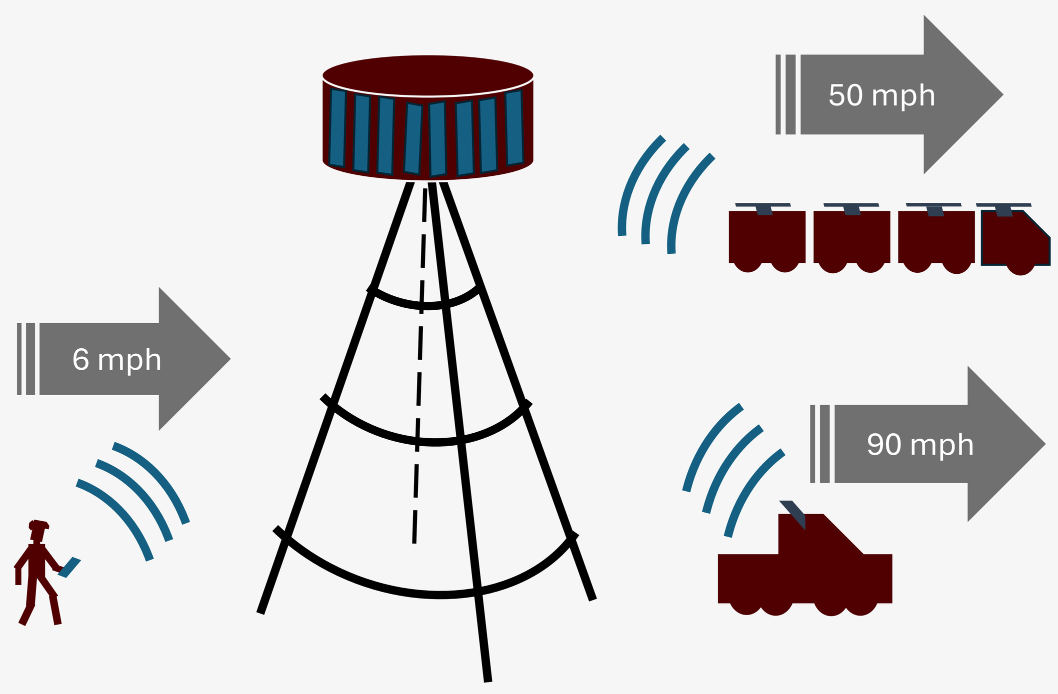
TEXAS A&M UNIVERSITY
Department of Electrical & Computer Engineering

Introduction



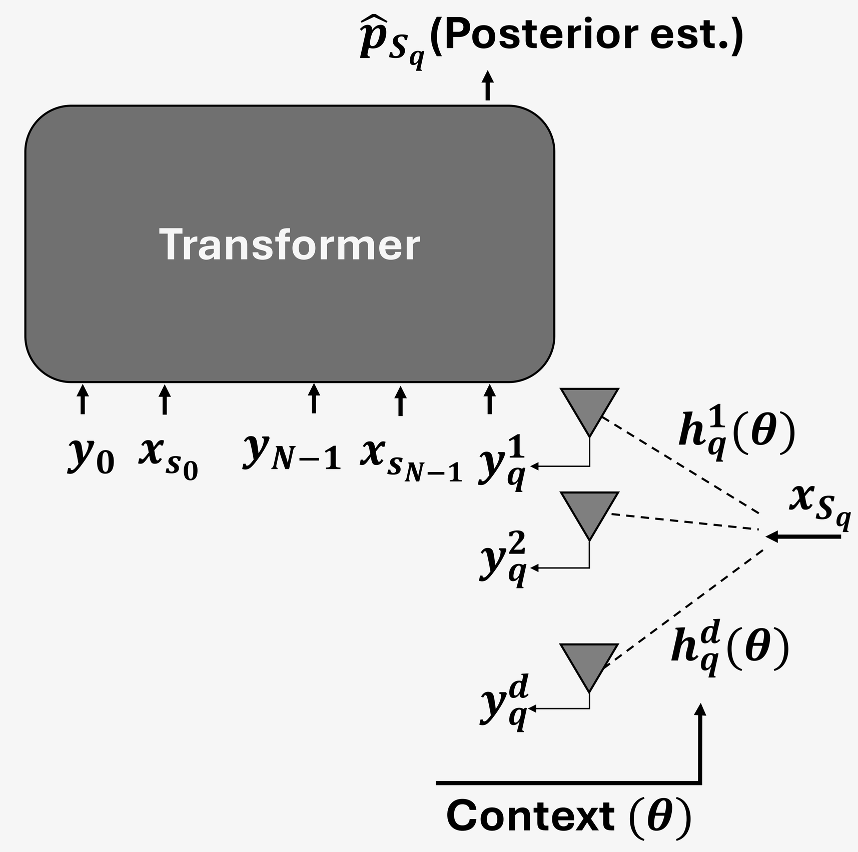
- **Transformers** excel at **in-context learning**
- Prompted with query-response pairs, it can generate **responses** by understanding the **context**
- A **single** transformer can handle various contexts

- Estimation problems in **wireless** can also be viewed as in-context learning problems
- **Context** is the scattering environment (antenna correlation), mobility, delay-spread, etc.
- **Pilots** are the in-context examples (prompts)



Methodology

- The **channel** is modeled using a **hierarchical** prior, where the **context** variable $\boldsymbol{\theta}$ is chosen first, and the channel is sampled based on the **conditional** distribution
- The context is the type of the channel under consideration



- The transformer is trained to estimate the **posterior** of the next symbol, provided a sequence of **received** symbols and **pilots**
- We show that transformers can perform **efficient in-context estimation** for wireless
- We prove that **asymptotically**, transformers learn the optimal estimator

Theoretical Results

- Consider a single layer **Softmax Attention Transformer (SAT)** to make a posterior estimate for S_q given $y_q = Hx_{S_q} + z_q$ and $y_{1:N}, x_{s_{1:N}}$
- Let $W = W_K^T W_Q$ be the attention matrix. The transformer estimate for the posterior probability of $S_q = i$ is given by

$$\hat{p}_i^{\mathrm{TF}}(y_q, y_{1:N}, s_{1:N}; W) = \frac{\sum_{n \in N_i(N)} \exp(y_q^T W y_n)}{\sum_{j \in [S]} \sum_{m \in N_j(N)} \exp(y_q^T W y_m)}$$

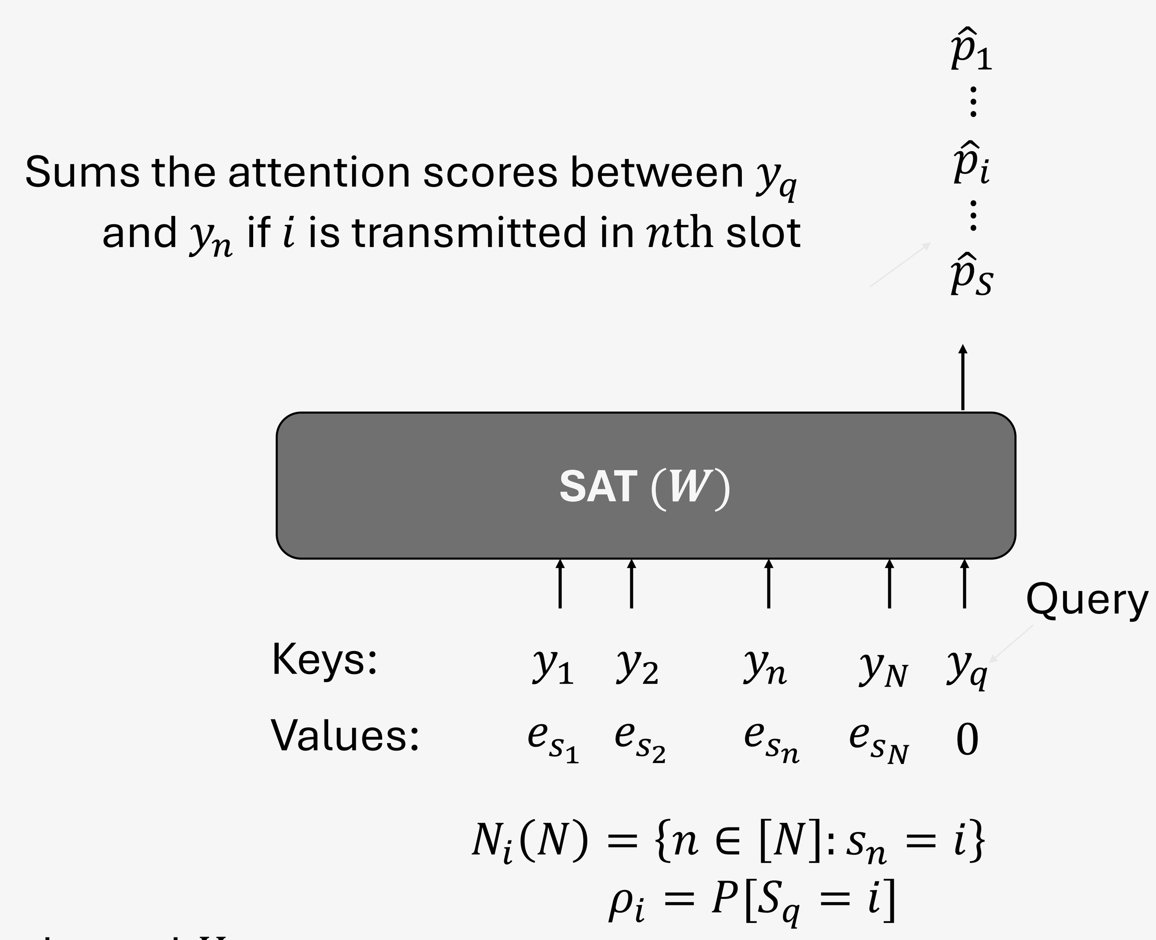
- By **law of large numbers**, we can show that

$$\lim_{N\to\infty} \hat{p}_i^{\mathrm{TF}}(y_q, y_{1:N}, s_{1:N}; W) = \frac{\rho_i \exp(y_q^T W H x_i)}{\sum_{j\in[S]} \rho_j \exp(y_q^T W H x_j)} = \hat{p}_i^{\mathrm{TF}}(y_q, H; W)$$

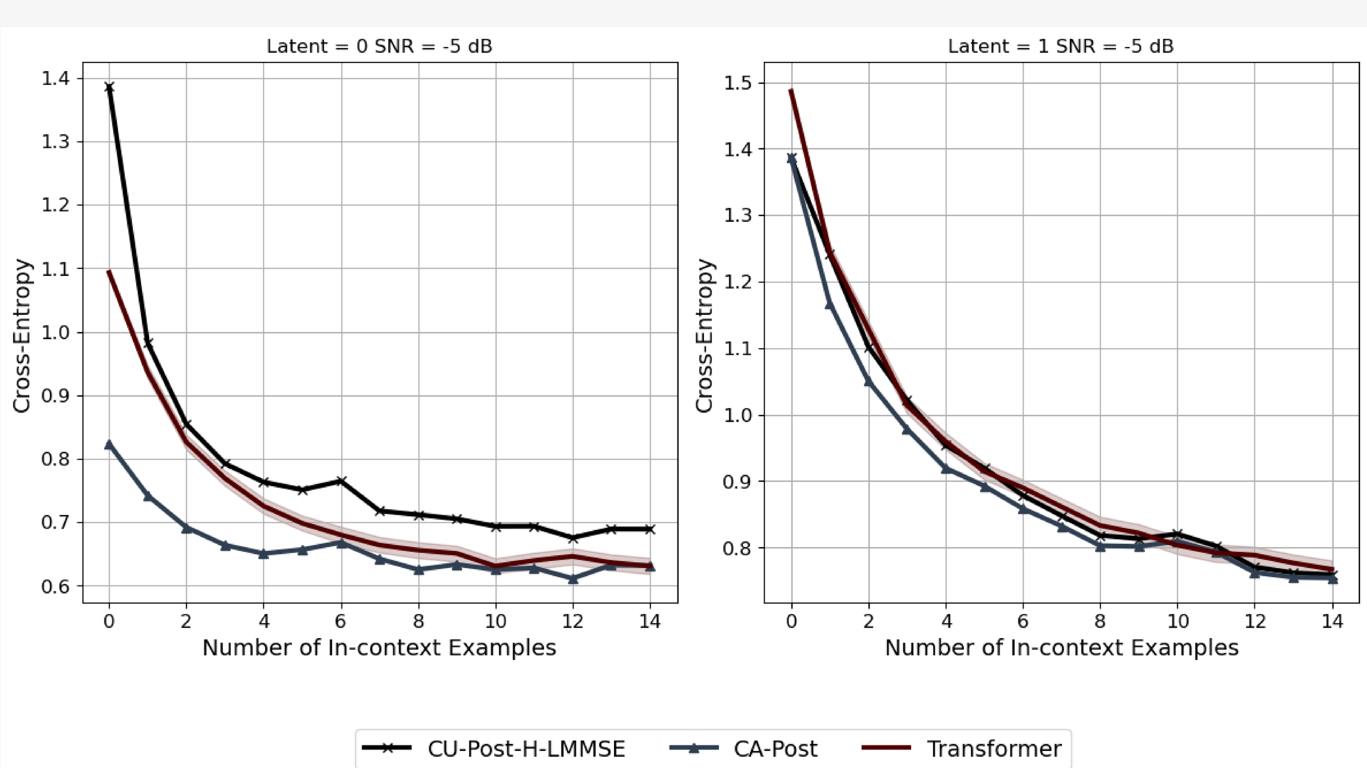
- If $||x|| = 1$, using $W = \Sigma^{-1}$ the inverse covariance of the noise, we get the optimal posterior **given** true channel H.
- The **population** cross-entropy loss at **long** prompt lengths is given by

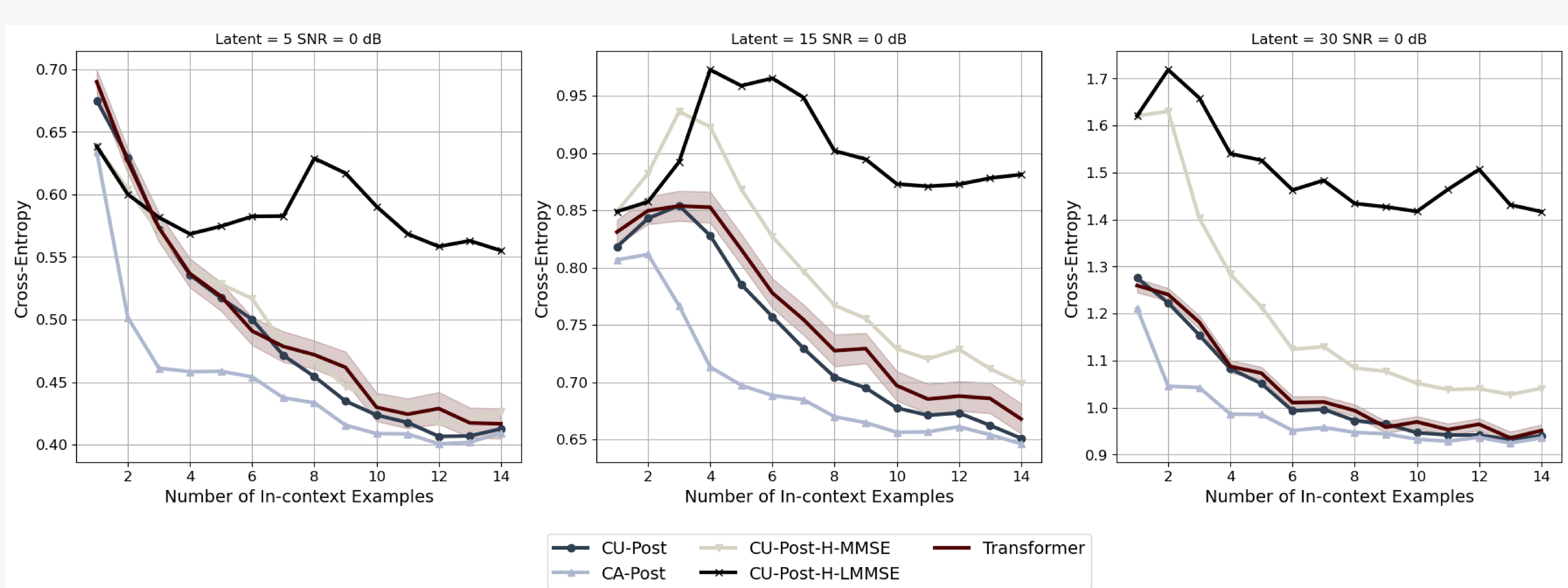
$$L(W) = -E\left[\sum_{j\in[S]} p_j(y_q, H; \Sigma^{-1}) \log \hat{p}_j^{\mathrm{TF}}(y_q, H; W)\right]$$

- L is **convex** in W with $W^* = \Sigma^{-1}$ as the global minimizer



Empirical Results





- Scenario 1: Context is the **Nature of Scattering** environment**.** One-ray model with a Line-of-Sight (LoS) vs fading environment with rich scattering respectively.
- Transformer performance approaches that of (context-aware) genie-aided LMMSE equalizer within few examples while outperforming the (context-unaware) LMMSE equalizer used in practice (black)

- Scenario 2: Context is **Mobility.** Each plot corresponds to the ground truth of $v = 5, 15, 30\ m/s$ respectively. The transformer performs as good as the computationally intensive Bayesian equalizer (dark blue) and approaches the (context-aware) genie-aided equalizer (bottom, light blue) within a few examples, while significantly outperforming the commonly used LMMSE equalizer (top, black), which suffers due to model mismatch.
- Metric: **Cross-entropy (CE):** Lesser CE implies higher quality LLRs giving better coding gains, if we employ the soft iterative decoders