Sample-Efficient Self-Interference Cancellation for In-Band Full Duplex Radios via In-Context Learning

Rushabha Balaji [*1] **Abhiram Kadiyala** [*1] **Danijela Cabric** [1] **Suhas Diggavi** [1]

Abstract

Digital self-interference cancellation is a central bottleneck in realising in-band full-duplex radios for 6G. Modern cancellers can model power-amplifier nonlinearities accurately, but their coefficients must be re-estimated from short calibration-interval pilots whenever the amplifier, front-end, or self-interference channel drifts. We recast this online calibration problem as in-context learning (ICL): a transformer pre-trained over a distribution of amplifier operating conditions, channel, and front-end conditions predicts the interference from a few calibration-interval examples supplied as context. On a synthetic model derived from real data, the proposed method matches the analytical Bayes-optimal predictor across context lengths. With real measured PA samples passed through a simulated Rician self-interference channel and I/Q imbalance, ICL is roughly $4.4\times$ more sample-efficient than a widely-linear alternating-least-squares baseline. On USRP N210 full-duplex captures, ICL reaches its NMSE floor in approximately $64\times$ fewer calibration samples than the strongest model-based baselines. When the context samples are appropriately normalized, we also show that ICL adapts zero-shot across nominally identical hardware units.

1. Introduction

Sixth-generation (6G) wireless networks are expected to deliver substantially higher spectral efficiency while natively supporting integrated sensing and communication over shared time–frequency resources (Liu et al., 2022; Roberts & Suraweera, 2022). In-band full-duplex (FD) communication, which enables a radio to transmit and receive simultaneously in the same frequency band, is a natural physical-layer mechanism for advancing both goals (Sabharwal et al., 2014; Duarte et al., 2012). By reusing the same time–frequency resource for transmission and reception, FD can improve spectral utilization and support sensing using the transmitted waveform. The main barrier to practical FD operation is self-interference (SI). Because the transmitter and receiver are co-located, the local transmit signal leaks into the receive chain and can exceed the desired received signal by tens of decibels (dBs). Consequently, SI must be suppressed to a sufficiently low level before reliable demodulation of the desired signal is possible (Kolodziej et al., 2019; Sabharwal et al., 2014).

Practical FD receivers suppress SI in stages (Bharadia et al., 2013; Kolodziej et al., 2019). An analog canceller is placed before the ADC. Since the transmit waveform is known, it constructs a delayed, phase-shifted, and scaled replica of the leakage and subtracts it at the receive front-end (Choi et al., 2010; Jain et al., 2011; Bharadia et al., 2013). This removes the dominant linear coupling path, so the residual no longer drives the ADC out of its dynamic range. If $y_{\mathrm{SI}}[n]$ denotes the total self-interference before digital cancellation, then after analog cancellation the residual received signal is

$$r[n] = y_{\mathrm{soi}}[n] + \underbrace{\big(y_{\mathrm{SI}}[n] - y_{\mathrm{AC}}[n]\big)}_{\text{residual SI}} + w[n], \qquad (1)$$

where $y_{\mathrm{AC}}[n]$ is the analog cancellation signal, $y_{\mathrm{soi}}[n]$ is the desired remote signal, and $w[n]$ denotes receiver noise. The role of digital SI cancellation is therefore to estimate and subtract the residual SI term. A substantial component of this residual arises from the signal-dependent nonlinear distortions in the transmit chain (Korpi et al., 2014b; Sabharwal et al., 2014; Ahmed & Eltawil, 2015). We therefore focus on modeling the PA-induced component of the residual SI (Morgan et al., 2006; Zhu et al., 2008). Let

$$\boldsymbol{x}_n = [x[n], x[n-1], \ldots, x[n-M]]^{\mathsf{T}}$$

denote a window of locally known transmit samples. Most conventional algorithms share a common structure. Specifically, after choosing a nonlinear feature representation of $\boldsymbol{x}_n$, the residual SI is modeled as a linear combination of these features. Let $\boldsymbol{\phi}\colon \mathbb{C}^{M+1} \to \mathbb{C}^D$ denote such a feature map, where D is the resulting feature dimension. Then the residual SI predictor can be written as

$$\hat{y}_{\mathrm{SI}}[n] = \boldsymbol{\phi}(\boldsymbol{x}_n)^{\mathsf{T}}\boldsymbol{a}, \qquad (2)$$

[*]Equal contribution [1]University of California, Los Angeles. Correspondence to: Rushabha Balaji <rubalaji99@g.ucla.edu>.

Preprint. May 3, 2026.

where $\boldsymbol{a} \in \mathbb{C}^D$ is a complex coefficient vector. Different classical cancellers correspond to different choices of $\boldsymbol{\phi}$; for example, memory polynomial (MP), generalized MP (GMP), or Hammerstein-/Wiener-type basis functions (Morgan et al., 2006; Kim & Konstantinou, 2001; Korpi et al., 2014a). More recent approaches replace or augment these hand-crafted bases with kernel and neural-network representations (Balatsoukas-Stimming, 2018; Kim et al., 2024; Wu et al., 2024). Thus, once the feature representation is fixed, digital SI cancellation reduces to estimating the coefficient vector $\boldsymbol{a}$ from calibration samples.

Given a calibration sample set $\mathcal{D}_K = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^K$ collected without the received signal present, the coefficients are estimated from these samples. In practice, these calibration samples can be obtained by internally terminating the receiver antenna. We refer to this timeframe as the *calibration interval* (Sabharwal et al., 2014; Kim et al., 2021). For a fixed $\boldsymbol{\phi}$, the coefficient estimation then reduces to ridge regression in the induced feature space. Let

$$\boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{\phi}(\boldsymbol{x}_1)^\mathsf{T} \\ \vdots \\ \boldsymbol{\phi}(\boldsymbol{x}_K)^\mathsf{T} \end{bmatrix} \in \mathbb{C}^{K \times D}, \qquad \boldsymbol{y} = [y_1, \ldots, y_K]^\mathsf{T}.$$

Then the regularized coefficient estimate is

$$\hat{\boldsymbol{a}} = \arg\min_{\boldsymbol{a} \in \mathbb{C}^D} \sum_{i=1}^{K} \left| y_i - \boldsymbol{\phi}(\boldsymbol{x}_i)^\mathsf{T} \boldsymbol{a} \right|^2 + \lambda \|\boldsymbol{a}\|_2^2 \tag{3}$$

$$= (\boldsymbol{\Phi}^\mathsf{H} \boldsymbol{\Phi} + \lambda \boldsymbol{I})^{-1} \boldsymbol{\Phi}^\mathsf{H} \boldsymbol{y}. \tag{4}$$

Here $\boldsymbol{\Phi}^\mathsf{H}$ denotes the Hermitian transpose of $\boldsymbol{\Phi}$. Other regularizers, for example, L_1 regression can be used for a sparser estimation (Wang et al., 2023). With sufficiently many calibration samples, this pipeline routinely achieves 40–50 dB of digital SI suppression (Korpi et al., 2014b; Kim et al., 2024). The difficulty is that the coefficients are not stationary. The model coefficients are sensitive to junction temperature, bias point, supply voltage changes, instantaneous output back-off, and component aging (Vuolevi & Rahkonen, 2003). Moreover, the SI channel between the transmit and receive chains can vary rapidly because of changes in the local environment, forcing faster drifts in the coefficients. The standard mitigation is to restart a calibration interval and refit $\boldsymbol{a}$ from a new $\mathcal{D}_K$ (Sabharwal et al., 2014; Kim et al., 2021; 2024; Balatsoukas-Stimming, 2018). This refit is expensive: high-order MP/GMP models with deep memory require many samples for $\boldsymbol{\Phi}^\mathsf{H}\boldsymbol{\Phi}$ to be well conditioned, and frequent calibration intervals reduce the effective spectral efficiency. The cost is amplified in dense 6G base-station deployments, where a single antenna panel may host hundreds of nominally identical PA chains, each with its own process variation and thermal conditions (Larsson et al., 2014). *The resulting challenge is to estimate an accurate SI predictor for each PA and its operating condition by using only a small number of calibration-interval samples.*

In-context learning (ICL) is well suited to this few-sample calibration problem. First observed in large language models, ICL refers to the ability of a pretrained sequence model to adapt to a new task in the forward pass by conditioning on a small set of input–output examples provided in its context (Brown et al., 2020). This behavior has been shown in the linear regression setting, where pre-trained transformers can infer the underlying linear map from examples (Garg et al., 2022; Zhang et al., 2024). More recently, related ICL capabilities have been demonstrated for nonlinear functionals and kernel-based regression as well (Cheng et al., 2024). From a Bayesian perspective, ICL acts as an approximate posterior predictor where the context identifies a latent task from a distribution learned during pre-training (Xie et al., 2022; Panwar et al., 2024; Akyürek et al., 2023; Raventós et al., 2023; Zhou et al., 2024). This framework is a natural fit for FD calibration. By conditioning on calibration-interval examples, a pre-trained transformer can potentially learn a more powerful feature mapping function to predict the SI than the fixed hand-crafted ones used before, all while using a smaller number of calibration points. We demonstrate the power of ICL using a modified GPT-2 style architecture (Radford et al., 2019) through extensive simulation and hardware experiments. Our main contributions are as follows.

1. We formulate digital SI cancellation as an ICL problem, where K calibration-interval transmit–receive pairs define the context and the model predicts the residual SI for a query transmit window.
2. On a synthetic benchmark derived from the OpenDPD dataset (Wu et al., 2024), in the tractable Gaussian coefficient-drift setting ICL matches the analytic Bayes-optimal predictor; in the intractable setting where a Rician SI channel and I/Q imbalance are applied on top of measured PA samples, ICL is $\approx 4.4\times$ more sample-efficient than the strongest classical baseline.
3. On over-the-air USRP N210 captures, a single pre-trained transformer outperforms classical baselines across operating points and transfers across nominally identical hardware units without retraining. By performing adaptation and implicit model selection in context, ICL offers a learned alternative to the conventional setup of maintaining separate coefficient lookup tables for each device, power level, and bandwidth setting.

2. System Model

Building on the post-analog-cancellation decomposition in (1), we focus on the digital cancellation stage. Without loss of generality, we treat the residual $y_{\mathrm{SI}}[n] - y_{\mathrm{AC}}[n]$

as the SI seen by the digital canceller and, with an abuse of notation, continue to denote it by $y_{\mathrm{SI}}[n]$ throughout the remainder of the paper. The ADC observation then admits the decomposition

$$r[n] = y_{\mathrm{SI}}[n] + y_{\mathrm{soi}}[n] + w[n]. \tag{5}$$

During the calibration, the signal of interest is absent, $y_{\mathrm{soi}}[n] = 0$, and the canceller observes $y_{\mathrm{SI}}[n] + w[n]$.

The residual SI is generated by a cascade of physical processes set by the radio architecture. Let M denote an upper bound on the joint memory of the transceiver and the SI channel. The SI signal admits the standard cascade decomposition

$$y_{\mathrm{SI}}[n] = f\big(h\big(\psi\big(x[n+M-1], \ldots, x[n], \\ x[n-1], \ldots, x[n-M]\big) + v[n]\big)\big), \tag{6}$$

where $\psi(\cdot)$ models the transmitter chain including its memory and nonlinear distortion, $h(\cdot)$ is the multipath SI channel, $f(\cdot)$ models the receiver chain, and $v[n]$ is the broadband transmitter noise emitted after the PA. We treat $f(\cdot)$ as linear, since the PA dominates the digital SI nonlinearity (Korpi et al., 2014b); any remaining residual is absorbed into $w[n]$.

A digital SI canceller produces an estimate $\hat{y}_{\mathrm{SI}}[n]$ of $y_{\mathrm{SI}}[n]$ and reports the suppression

$$C \;=\; 10\log_{10}\left(\frac{\mathbb{E}\big[\,|y_{\mathrm{SI}}[n]|^2\big]}{\mathbb{E}\big[\,|y_{\mathrm{SI}}[n] - \hat{y}_{\mathrm{SI}}[n]|^2\big]}\right), \tag{7}$$

or, equivalently, the normalized mean-squared error $\mathrm{NMSE} = -C$.

3. Classical Estimation and Bayesian View

We begin with the simplest case of a unit-tap SI channel, $h(\cdot) \equiv 1$, so that (6) reduces to a single nonlinear, finite-memory transmitter response; the multi-tap Rician channel is reintroduced later in this section. For simplicity, we adopt the standard odd-order memory polynomial (MP) to model $\psi(\cdot)$ (Morgan et al., 2006; Kim & Konstantinou, 2001) but any of the previous feature maps could be used.

$$\begin{aligned} y_{\mathrm{SI}}[n] &= \sum_{p=0}^{P}\sum_{m=0}^{M} a_{p,m}\, x[n-m]|x[n-m]|^{2p} \\ &= \boldsymbol{\phi}_{\mathrm{MP},n}^{\mathsf{T}}\boldsymbol{a}, \end{aligned} \tag{8}$$

where P indexes the odd-order nonlinearity terms $|x|^{2p}$ for $p = 0, 1, \ldots, P$. The coefficient vector $\boldsymbol{a} \in \mathbb{C}^{(P+1)(M+1)}$. Operationally, an offline training dataset $\mathcal{D}_{\mathrm{train}}$ is collected once per radio and a model-order sweep over (P, M) selects $(\hat{P}, \hat{M})$ minimizing hold-out NMSE on $\mathcal{D}_{\mathrm{train}}$. The required memory depth M is set primarily by the signal bandwidth and the frequency response of the matching network, while the polynomial order P tracks how deep the PA is driven into compression.

The fitted coefficients $\boldsymbol{a}$ then drift between the offline fit and the deployment calibration interval. The standard adaptive remedy is to recompute the regularised pseudoinverse of (4) on the growing calibration window:

$$\hat{\boldsymbol{a}}_n = \boldsymbol{a}^\star + \big(\boldsymbol{\Phi}_n^{\mathsf{H}}\boldsymbol{\Phi}_n + \lambda \boldsymbol{I}\big)^{-1}\boldsymbol{\Phi}_n^{\mathsf{H}}(\boldsymbol{y}_n - \boldsymbol{\Phi}_n\boldsymbol{a}^\star), \tag{9}$$

which warm-starts from the offline baseline $\boldsymbol{a}^\star$ and shrinks toward it with strength λ. Recursive least squares (RLS) (Kay, 1993) is the streaming variant of (9), and has been applied to digital SIC in full-duplex transceivers (Korpi et al., 2014a; Kim et al., 2024). A comprehensive way to handle drift is to model it as $\boldsymbol{a} = \boldsymbol{a}^\star + \boldsymbol{\varepsilon}$, with $\boldsymbol{a}^\star$ the offline baseline and $\boldsymbol{\varepsilon}$ a drift random variable, and to seek the MAP estimate

$$\hat{\boldsymbol{a}}_{\mathrm{MAP}} = \arg\max_{\boldsymbol{a}}\; p(\boldsymbol{a} \mid \mathcal{D}_K). \tag{10}$$

Example: Gaussian perturbation. If the perturbation is Gaussian, $\boldsymbol{\varepsilon} \sim \mathcal{CN}(\boldsymbol{0}, \boldsymbol{\Sigma}_a)$, and the receiver noise has variance σ_w^2, then the posterior over $\boldsymbol{a}$ is Gaussian and the MAP, MMSE, and posterior-mean estimators all coincide (Kay, 1993). Throughout this paper we use *Bayes-optimal* to refer specifically to this Bayesian MMSE estimator under the true prior. The Bayes-optimal predictor for a query feature $\boldsymbol{\phi}_q$ and its mean-squared error are

$$\hat{y}_{\mathrm{Bayes}} = \boldsymbol{\phi}_q^{\mathsf{T}}\big(\boldsymbol{a}^\star + \boldsymbol{\Sigma}_a\boldsymbol{\Phi}^{\mathsf{H}}(\boldsymbol{\Phi}\boldsymbol{\Sigma}_a\boldsymbol{\Phi}^{\mathsf{H}} + \sigma_w^2\boldsymbol{I})^{-1}(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{a}^\star)\big), \tag{11}$$

$$\mathrm{MSE}_{\mathrm{Bayes}} = \sigma_w^2 + \boldsymbol{\phi}_q^{\mathsf{T}}\big(\boldsymbol{\Sigma}_a - \boldsymbol{\Sigma}_a\boldsymbol{\Phi}^{\mathsf{H}}(\boldsymbol{\Phi}\boldsymbol{\Sigma}_a\boldsymbol{\Phi}^{\mathsf{H}} + \sigma_w^2\boldsymbol{I})^{-1}\boldsymbol{\Phi}\boldsymbol{\Sigma}_a\big)\boldsymbol{\phi}_q^*. \tag{12}$$

Multipath channel. Reintroducing the multipath SI channel $h(\cdot)$ in (6) as a tap vector $\boldsymbol{h}$ couples the nonlinear PA output with the channel response, so RLS on the PA basis alone is no longer matched to the data-generating process. A classical replacement is alternating least squares (ALS) (Kolda & Bader, 2009), which exploits the bilinear structure of the PA–channel cascade: with the PA coefficients held fixed, the ALS algorithm solves a linear least-squares problem in $\boldsymbol{h}$, and with $\boldsymbol{h}$ fixed it solves another in $\boldsymbol{a}$, alternating until convergence. Joint PA and SI-channel estimation has also been studied with related techniques in full-duplex SIC (Ahmed & Eltawil, 2015; Korpi et al., 2015). Once additional impairments are factored in, such as I/Q imbalance, finite-sample drift correlations, and non-Gaussian process noise, the Bayes-optimal estimator is no longer analytically tractable, and each new impairment requires a corresponding increase in algorithmic complexity

to maintain performance. ICL replaces this growing per-impairment complexity with a single pre-trained model. We develop this method in the next section.

4. Proposed Methodology

We treat a calibration-interval batch $\mathcal{D}_K = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^K$ as the *context* for a sequence model and ask the model to predict the residual SI sample $\hat{y}_q$ for a query $\boldsymbol{x}_q$ under the same operating conditions as the calibration interval. The model is pre-trained over a distribution of PA coefficients, SI channel, and hardware impairments. Inference is a single forward pass conditioned on the K context pairs. We adopt the GPT-style transformer used by Garg et al. (2022) for in-context regression and modify three components to match the FD calibration setting: the tokenization, the attention pattern, and a per-context gain normalization.

Tokenization. For the i-th context example, we take the transmit window $\boldsymbol{x}_i \in \mathbb{C}^{M+1}$ and the scalar receive sample $y_i \in \mathbb{C}$ from the calibration set $\mathcal{D}_K$. We stack their real and imaginary parts into a real-valued token

$$\boldsymbol{z}_i = \big[\Re(\boldsymbol{x}_i),\, \Im(\boldsymbol{x}_i),\, \Re(y_i),\, \Im(y_i)\big]^{\mathsf{T}} \in \mathbb{R}^{2M+4}, \quad (13)$$

The query token $\boldsymbol{z}_q$ is constructed in the same way, with its two coordinates corresponding to y_q set to zero. The token sequence $(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_K, \boldsymbol{z}_q)$ is linearly embedded and processed by the transformer. The final query-token representation is passed through an MLP head to produce two outputs, corresponding to $\Re(\hat{y}_q)$ and $\Im(\hat{y}_q)$.

The construction above fixes how each example is embedded but not how the K context examples themselves are drawn from a capture. In practice, the transmitted signal can sometimes occupy only a fraction $\beta \in (0, 1]$ of the ADC sampling bandwidth rather than the full Nyquist band. At narrowband settings ($\beta \ll 1$) consecutive ADC samples are heavily oversampled, so windows taken at unit stride produce near-duplicate $\boldsymbol{x}_i, \boldsymbol{x}_{i+1}$. For the classical estimator of (4) this is fatal: neighboring rows of $\boldsymbol{\Phi}$ become collinear, the Gram matrix $\boldsymbol{\Phi}^{\mathsf{H}}\boldsymbol{\Phi}$ is ill-conditioned, and the ridge solution collapses toward simply copying the most recent $(\boldsymbol{x}_i, y_i)$ pair. The transformer is not immune: redundant context tokens shrink the effective rank of the in-context regression problem. We therefore apply a stride $d = \lceil 1/\beta \rceil$ between the center taps of successive context windows, so that the inter-window spacing is approximately one signal-Nyquist period. Section G studies the ablation of this data pre-processing technique.

Attention pattern and positional encoding. During the calibration interval, the context examples are drawn from the same stationary radio state. Consequently, the joint distribution of $(\boldsymbol{z}_1, \ldots, \boldsymbol{z}_K)$ is exchangeable, and a well-specified predictor of y_q should be invariant under permutations of the context. Two architectural choices encode this prior. First, we omit positional encodings altogether, so reordering the context cannot change the prediction. Second, we replace the standard causal attention mask with a *prefix* attention mask (Raffel et al., 2020). Under this mask the K context tokens attend to one another bidirectionally: each $\boldsymbol{z}_i$ sees every other $\boldsymbol{z}_j$, not just the ones before it. The query token $\boldsymbol{z}_q$ is appended to the prefix and attends to all K context tokens at once. The two halves of this mask mirror the two terms in the closed-form least-squares solution (4): the bidirectional context-to-context attention plays the role of the symmetric Gram operator $\boldsymbol{\Phi}^{\mathsf{H}}\boldsymbol{\Phi}$, and the query-to-context attention plays the role of $\boldsymbol{\Phi}^{\mathsf{H}}\boldsymbol{y}$. Similar connections have been made in the *mesa-layer* of von Oswald et al. (2023).

Gain normalization. The absolute scale and phase of the calibration samples y_i vary substantially across hardware units due to LNA gain, AGC settings, and mixer phase. These variables are not part of the nonlinear PA structure that we want the transformer to model, but are inherently part of the calibration samples. If not removed, they would force the network to learn per-device gain patterns. To remedy this problem, we normalize y_i with a single complex scalar gain $\hat{\alpha}$ per context. Let $x_i \triangleq x[i]$ denote the center-tap transmit sample of the i-th context window, collected into the vector $\boldsymbol{x}_c = [x_1, \ldots, x_K]^{\mathsf{T}} \in \mathbb{C}^K$. The gain estimate $\hat{\alpha}$ is given as,

$$\hat{\alpha} \;=\; \frac{\boldsymbol{x}_c^{\mathsf{H}}\boldsymbol{y}}{\boldsymbol{x}_c^{\mathsf{H}}\boldsymbol{x}_c + \epsilon} \;=\; \frac{\sum_{i=1}^K x_i^* \, y_i}{\sum_{i=1}^K |x_i|^2 + \epsilon}, \quad (14)$$

where $\epsilon > 0$ is a small numerical regularizer; if $|\hat{\alpha}|$ falls below a fixed floor (indicating numerical underflow) we replace it with $\hat{\alpha} = 1$. Each of the K context receive samples is then divided by $\hat{\alpha}$ before tokenization, so the transformer sees the normalized pairs $\{(\boldsymbol{x}_i,\, y_i/\hat{\alpha})\}_{i=1}^K$. It produces a normalized query prediction $\hat{y}_{\mathrm{norm}}$, which we multiply back by $\hat{\alpha}$ to recover the final SI estimate $\hat{y}_q = \hat{\alpha}\,\hat{y}_{\mathrm{norm}}$. Because $\hat{\alpha}$ is a function of the context only and is computed without access to y_q, this enables cross-device generalization. It absorbs the linear, gain-only component of the transmit-to-receive map and leaves the nonlinear and memory structure for the transformer to learn. Section 5.2 shows that disabling (14) costs roughly 10 dB of cross-device NMSE.

Training objective. With θ denoting the network parameters and $e \in \{\Re(y_q - \hat{y}_q),\, \Im(y_q - \hat{y}_q)\}$ the per-component query error, we train with the *reverse pseudo-Huber* loss (Odland, 2024)

$$\mathcal{L}(\theta) = \mathbb{E}[\rho_\delta(e)], \quad \rho_\delta(e) = \tfrac{\delta^2}{2}\big(u\sqrt{u^2+1} + \sinh^{-1} u\big), \quad (15)$$

with $u \triangleq |e|/\delta$ and a fixed scale $\delta > 0$. Unlike the standard (pseudo-)Huber, the gradient of ρ_δ has constant magnitude

near zero and grows linearly in the tails, so the loss behaves like L_1 near zero and L_2 for large errors. The L_1-like behavior on small residuals is what we need at deep NMSE: an MSE objective vanishes quadratically here and supplies almost no gradient on the residuals that dominate this regime. More details regarding the loss function are in Section A.

5. Results

This section reports both synthetic and hardware results. We probe three regimes of increasing realism – a tractable PA-only setting, a simulated FD link with multipath SI and I/Q imbalance, and over-the-air USRP N210 captures. We benchmark ICL against fair, $\boldsymbol{a}^\star$-informed classical baselines. We do not use an analog canceller, so the digital stage is responsible for suppressing both the linear SI path and the PA-induced nonlinearities. This is consistent with prior captures-only studies that report digital-only suppression on SDR testbeds, including the neural-network canceller of Balatsoukas-Stimming (2018) and the widely-linear USRP measurement of Wahab et al. (2022). For the hardware captures an external RF attenuator is inserted in the direct path to keep receive samples below the ADC full-scale; capture details are in Section D.

5.1. Simulation Results

ICL learns the Bayes-optimal solution. We first build a synthetic dataset whose ground-truth nonlinearity is calibrated to a real device. Using the OpenDPD DPA-200 MHz dataset (Wu et al., 2024), we sweep model orders to select the (P, M) pair with the lowest held-out NMSE. The resulting coefficient vector $\boldsymbol{a}^\star$ and order $(P^\star, M^\star)$ define our reference PA. The synthetic dataset is then generated by transmitting i.i.d. complex-Gaussian waveforms through this reference PA, with the coefficients sampled as $\boldsymbol{a} = \boldsymbol{a}^\star + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim \mathcal{CN}(\mathbf{0}, \boldsymbol{\Sigma}_a)$. Crucially, all K context examples and the query within a single sequence share the same perturbation $\boldsymbol{\varepsilon}$, so each sequence is one task and the transformer must infer it from context. Additive white Gaussian noise at a fixed SNR is added to the receive samples. The closed-form Bayes-optimal predictor and its MSE are given by Equations (11) and (12). Full dataset and training hyperparameters are in Sections A and B.

On in-distribution complex-Gaussian queries, Figure 1a shows that ICL tracks the analytic Bayes-optimal predictor to within 1 dB across the entire range $K \in [1, 128]$. The RLS-MP baseline, with ridge λ tuned, sits 0.5–2 dB above the bound. The gap reflects that the Bayes-optimal estimator (and ICL, which approximates it) exploits the per-coefficient prior covariance $\boldsymbol{\Sigma}_a$. This shows that the transformer has internalized the prior $p(\boldsymbol{\varepsilon})$ during pre-training and applies it as an effective estimator in context. Because pre-training uses complex-Gaussian samples, it is natural to ask whether ICL is merely memorizing the input statistics; to test this we evaluate the same checkpoint on out-of-distribution (OOD) OFDM queries with sub-carrier spacing matched to the DPA-200 MHz bandwidth, leaving the PA model and perturbation distribution unchanged. Figure 1b shows that the ICL curves track the analytic Bayes-optimal curve on OFDM input as well.

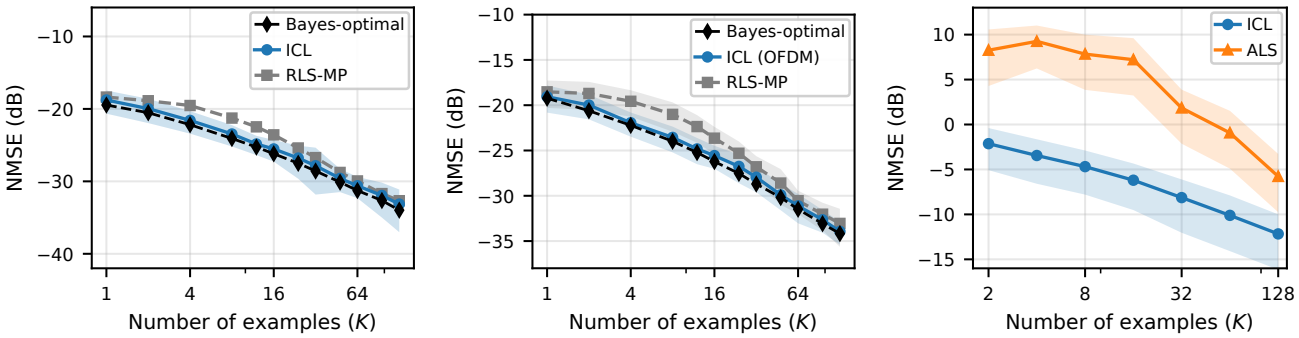
Beyond the tractable setting We now reintroduce the two impairments that make a real FD link harder than the PA-only setting: a multipath self-interference channel and front-end I/Q imbalance. To stress-test against PA model mismatch, we feed measured DPA-200 MHz IQ samples (held-out OpenDPD split) through an independent Rician multipath channel and widely-linear I/Q imbalance setting $y_{\text{IQ}} = c_1 y + c_2 y^*$; full details are in Section B. The joint posterior over $(\boldsymbol{a}, \boldsymbol{h}, c_1, c_2)$ is no longer tractable. We therefore benchmark against an $\boldsymbol{a}^\star$-informed ALS, with a widely-linear basis to absorb the I/Q imbalance and the same coefficient-drift prior (Section C). Figure 1c compares ICL with this ALS variant: ICL leads by 6–13 dB across $K \in [2, 128]$, and WL-ALS only matches ICL's $K{=}128$ NMSE at $K \approx 567$. This is consistent with the Bayesian view of Section 3: ICL behaves as an amortized approximation to the (intractable) Bayes-optimal estimator that ALS only reaches with many more samples.

5.2. Hardware Results

Setup. We validate ICL on over-the-air USRP N210 captures at 915 MHz, 25 MS/s, with a 30 dB attenuator inline between the transmit port and the transmit antenna[1]. The 915 MHz carrier is chosen for low band activity, and at this frequency the SI channel is single-tap, so coefficient drift is driven by the PA operating point alone — a simple proof-of-concept testbed for ICL. Three nominally identical N210 units are used: Device 1 supplies all training captures and devices 2–3 are held out for zero-shot cross-device tests. The training mix uses complex-Gaussian baseband waveforms with signal bandwidths spanning 2.5–20 MHz and transmit scales from 0.20 to 0.70, covering linear through soft-compression PA regimes. Figure 2 reports a single in-distribution evaluation cell at 16.25 MHz and transmit scale 0.60. Full setup is in Section D.

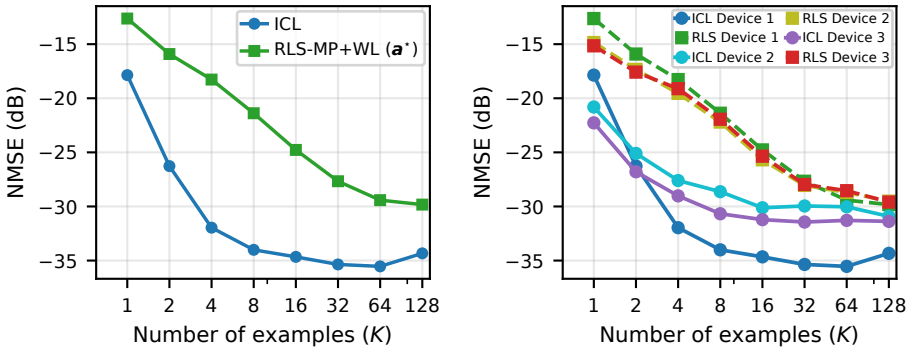
Same-device. On device 1 we report two MP+WL baselines so that the comparison is fair on both ends of the calibration-budget axis: a single fixed classical model order would either under-fit at large K or be ill-conditioned at small K, so reporting only one would handicap the classical estimator at the operating point we care about most. A *large* MP+WL model ($D{\approx}1100$) is the strongest classical choice

[1]This prevents the ADC from saturating and will not be required if analog SI cancellation is present.



(a) Gaussian coefficient drift. *(b)* OFDM-waveform queries. *(c)* Rician SI channel and I/Q imbalance.

Figure 1. **Synthetic dataset.** NMSE vs. K (lower is better); classical baselines warm-started at $\boldsymbol{a}^\star$. **(a)** Single-tap channel, Gaussian coefficient drift; RLS-MP ridge λ tuned by held-out CV. **(b)** OOD OFDM queries. **(c)** Real measured DPA-200 MHz IQ samples through a simulated Rician multipath channel and wideband I/Q imbalance; baseline is $\boldsymbol{a}^*$-informed widely-linear ALS.



(a) Same-device, Device 1. *(b)* Cross-device, Devices 1–3.

Figure 2. **Hardware results: USRP N210, 915 MHz, 25 MS/s.** NMSE vs. K (lower is better) at signal bandwidth 16.25 MHz, transmit scale 0.60. **(a)** Device 1 (training); RLS-MP+WL warm-started with the per-cell asymptotic $\boldsymbol{a}^\star$. **(b)** Same Device-1 checkpoint and warm-start, zero-shot on Devices 1–3; ICL solid, classical reference dashed.

at large K and catches up to ICL's NMSE floor only at $K \approx 512$; below that K, the model is under-determined and the offline anchor $\boldsymbol{a}^\star$ alone is too weak a prior. A *smaller*, hand-tuned MP+WL (D=751), with model order (P, M) fine-tuned for this operating condition using the appropriate data points from the training dataset, is a strong classical choice for the small-K regime that ICL operates in. This smaller baseline plateaus 4 dB above ICL's floor, while ICL itself is within 1 dB of that floor by $K \approx 8$. ICL is therefore an attractive option: it outperforms in the small-K regime, where the classical estimator needs hand-tuned priors to stay well-conditioned, and it remains more sample-efficient than the larger MP+WL model that performs well once a long calibration window is available.

Cross-device. Figure 2b shows performance of the model trained on device-1 applied zero-shot to devices 2 and 3. It converges within 8 context examples. Concretely, the gap to the classical baseline shrinks for devices 2 and 3. This transfer is enabled by the per-window gain normalization $\hat{\alpha}$ of (14), which factors out per-device linear variations so the transformer can re-use its learned nonlinear PA structure across devices; disabling $\hat{\alpha}$ flattens the K-dependence on device 3 almost entirely (Section J). Per-cell generalization across signal bandwidths and transmit scales is shown in Sections H and I.

6. Conclusion

We introduced an ICL formulation for full-duplex SI cancellation in which calibration pilots become context examples and a pre-trained transformer predicts the residual SI. The method recovers Bayes-optimal behavior under Gaussian PA-coefficient drift, improves over classical adaptive estimators under simulated channel and I/Q impairments, and adapts across hardware operating points on USRP N210 captures. A single pre-trained transformer thus replaces per-device coefficient tables and recurring ridge refits with on-the-fly few-shot calibration — collapsing a cost that scaled with every device and operating point into one that scales only with the in-context pilot count, a substitution whose value compounds at multi-chain 6G deployment scale.

References

Ahmed, E. and Eltawil, A. M. All-digital self-interference cancellation technique for full-duplex systems. *IEEE Transactions on Wireless Communications*, 14(7):3519–3532, 2015.

Akyürek, E., Schuurmans, D., Andreas, J., Ma, T., and Zhou, D. What learning algorithm is in-context learning? Investigations with linear models. In *International Conference on Learning Representations (ICLR)*, 2023.

Balatsoukas-Stimming, A. Non-linear digital self-interference cancellation for in-band full-duplex radios using neural networks. In *IEEE 19th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, pp. 1–5, 2018.

Bharadia, D., McMilin, E., and Katti, S. Full duplex radios. In *Proceedings of the ACM SIGCOMM 2013 Conference*, pp. 375–386, Hong Kong, China, 2013. ACM.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Cheng, X., Chen, Y., and Sra, S. Transformers implement functional gradient descent to learn non-linear functions in context. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2024.

Choi, J. I., Jain, M., Srinivasan, K., Levis, P., and Katti, S. Achieving single channel, full duplex wireless communication. In *Proceedings of the 16th Annual International Conference on Mobile Computing and Networking (MobiCom)*, pp. 1–12, Chicago, IL, USA, 2010. ACM.

Duarte, M., Dick, C., and Sabharwal, A. Experiment-driven characterization of full-duplex wireless systems. *IEEE Transactions on Wireless Communications*, 11(12):4296–4307, 2012.

Garg, S., Tsipras, D., Liang, P., and Valiant, G. What can transformers learn in-context? a case study of simple function classes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Jain, M., Choi, J. I., Kim, T., Bharadia, D., Seth, S., Srinivasan, K., Levis, P., Katti, S., and Sinha, P. Practical, real-time, full duplex wireless. In *Proceedings of the 17th Annual International Conference on Mobile Computing and Networking (MobiCom)*, pp. 301–312, Las Vegas, NV, USA, 2011. ACM.

Kay, S. M. *Fundamentals of Statistical Signal Processing: Estimation Theory*. Prentice Hall, 1993.

Kim, J. and Konstantinou, K. Digital predistortion of wideband signals based on power amplifier model with memory. *Electronics Letters*, 37(23):1417–1418, 2001.

Kim, J., Lee, H., Do, H., Choi, J., Park, J., Shin, W., Eldar, Y. C., and Lee, N. On the learning of digital self-interference cancellation in full-duplex radios. *IEEE Wireless Communications*, 2024. Also available as arXiv:2308.05966.

Kim, T., Min, K., and Park, S. Self-interference channel training for full-duplex massive MIMO systems. *Sensors*, 21(9):3250, 2021. doi: 10.3390/s21093250.

Kolda, T. G. and Bader, B. W. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.

Kolodziej, K. E., Perry, B. T., and Herd, J. S. In-band full-duplex technology: Techniques and systems survey. *IEEE Transactions on Microwave Theory and Techniques*, 67 (7):3025–3041, 2019.

Korpi, D., Anttila, L., Syrjälä, V., and Valkama, M. Widely linear digital self-interference cancellation in direct-conversion full-duplex transceiver. *IEEE Journal on Selected Areas in Communications*, 32(9):1674–1687, 2014a.

Korpi, D., Riihonen, T., Syrjälä, V., Anttila, L., Valkama, M., and Wichman, R. Full-duplex transceiver system calculations: Analysis of ADC and linearity challenges. *IEEE Transactions on Wireless Communications*, 13(7): 3821–3836, 2014b.

Korpi, D., Choi, Y.-S., Huusari, T., Anttila, L., Talwar, S., and Valkama, M. Adaptive nonlinear digital self-interference cancellation for mobile inband full-duplex radio: Algorithms and RF measurements. In *IEEE Global Communications Conference (GLOBECOM)*, 2015.

Larsson, E. G., Edfors, O., Tufvesson, F., and Marzetta, T. L. Massive MIMO for next generation wireless systems. *IEEE Communications Magazine*, 52(2):186–195, 2014. doi: 10.1109/MCOM.2014.6736761.

Liu, F., Cui, Y., Masouros, C., Xu, J., Han, T. X., Eldar, Y. C., and Buzzi, S. Integrated sensing and communications: Toward dual-functional wireless networks for 6G and beyond. *IEEE Journal on Selected Areas in Communications*, 40(6):1728–1767, 2022. doi: 10.1109/JSAC.2022.3156632.

Morgan, D. R., Ma, Z., Kim, J., Zierdt, M. G., and Pastalan, J. A generalized memory polynomial model for digital predistortion of RF power amplifiers. *IEEE Transactions on Signal Processing*, 54(10):3852–3860, 2006.

Odland, T. The reverse pseudo-Huber loss function. https://tommyodland.com/articles/2024/the-reverse-pseudo-huber-loss-function/, 2024. Blog post.

Panwar, M., Ahuja, K., and Goyal, N. In-context learning through the Bayesian prism. In *International Conference on Learning Representations (ICLR)*, 2024. arXiv:2306.04891.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., and Sutskever, I. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019. GPT-2 technical report.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., and Liu, P. J. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21 (140):1–67, 2020.

Raventós, A., Paul, M., Chen, F., and Ganguli, S. Pretraining task diversity and the emergence of non-Bayesian in-context learning for regression. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Roberts, I. P. and Suraweera, H. A. Full-duplex transceivers for next-generation wireless communication systems. In *Fundamentals of 6G Communications and Networking*. Springer, 2022.

Sabharwal, A., Schniter, P., Guo, D., Bliss, D. W., Rangarajan, S., and Wichman, R. In-band full-duplex wireless: Challenges and opportunities. *IEEE Journal on Selected Areas in Communications*, 32(9):1637–1652, 2014.

von Oswald, J., Niklasson, E., Schlegel, M., Kobayashi, S., Zucchet, N., Scherrer, N., Miller, N., Sandler, M., Arcas, B. A. y., Vladymyrov, M., Pascanu, R., and Sacramento, J. Uncovering mesa-optimization algorithms in transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Vuolevi, J. and Rahkonen, T. *Distortion in RF Power Amplifiers*. Artech House, Norwood, MA, 2003. ISBN 978-1-58053-539-7.

Wahab, A., Kim, Y.-H., and Maqsood, M. Widely-linear digital self-interference cancellation in full-duplex USRP transceiver. *Sensors*, 22(24):9607, 2022.

Wang, D., Lei, Y., and Zeng, L. A pruning method of the generalized memory polynomial model for power amplifiers based on the LASSO regression. In *IEEE 6th International Conference on Electronic Information and Communication Technology (ICEICT)*, pp. 182–185, Qingdao, China, 2023. doi: 10.1109/ICEICT57916.2023.10245788.

Wu, Y., Singh, G. D., Beikmirza, M., de Vreede, L. C. N., Alavi, M., and Gao, C. OpenDPD: An open-source end-to-end learning and benchmarking framework for wideband power amplifier modeling and digital pre-distortion. In *IEEE International Symposium on Circuits and Systems (ISCAS)*, 2024. arXiv:2401.08318.

Xie, S. M., Raghunathan, A., Liang, P., and Ma, T. An explanation of in-context learning as implicit Bayesian inference. In *International Conference on Learning Representations (ICLR)*, 2022.

Zhang, R., Frei, S., and Bartlett, P. L. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 2024. Also available as arXiv:2306.09927.

Zhou, R., Tian, C., and Diggavi, S. Transformers learn variable-order Markov chains in-context. *arXiv preprint arXiv:2410.05493*, 2024.

Zhu, A., Pedro, J. C., and Brazil, T. F. Dynamic deviation reduction-based Volterra behavioral modeling of RF power amplifiers. *IEEE Transactions on Microwave Theory and Techniques*, 54(12):4323–4332, 2008.

A. Training Configuration

A.1. Architecture and Optimisation

Table 1 lists the architecture, optimisation, and data-pipeline hyperparameters used for every checkpoint reported in the paper.

Table 1. Training hyperparameters.

Architecture	
Layers / heads / d_{model} / d_{ff}	12 / 4 / 256 / 1024
Token / output dimension	30 / 2
Attention mask	prefix (Raffel et al. 2020)
Positional encoding	none
Optimisation	
Optimiser	Adam, $\beta = (0.9, 0.999)$
Learning rate (peak / floor)	10^{-4} / 10^{-6}
Schedule	linear warm-up 2k steps, cosine decay
Batch size / total steps	1024 / 350k
Loss	reverse pseudo-Huber, δ=1
Context-length curriculum (Garg et al. 2022)	
$K_{\max}$	128
Ramp phase (0 – 175k steps)	$K \sim \mathcal{U}(1, K_{\max}(t))$, $K_{\max}(t)$ linear $1 \to K_{\max}$
Uniform phase (175k – 262.5k)	$K \sim \mathcal{U}(1, K_{\max})$
Long-context hold (262.5k – 350k)	$K \sim \mathcal{U}(K_{\max}/2, K_{\max})$

We follow the linear ramp curriculum of Garg et al. (2022) for the first 175k steps. Two modifications stabilize training at the operating points that matter for FD-SIC. First, after the ramp we add a *long-context hold* for the final 25% of training, in which K is drawn from $\mathcal{U}(K_{\max}/2, K_{\max})$; this reverses the bias toward small K that the linear ramp would otherwise leave behind, and is what lets the model push past the saturation plateau visible at moderate K in early checkpoints. Second, for the hardware checkpoint we further inject *short-context bursts* in the post-ramp phases by drawing $K \sim \mathcal{U}(2, 4)$ with probability p_{burst}=0.1; this exposes the per-context gain estimator $\hat{\alpha}$ in (14) to the regime where it is ill-conditioned (only a few transmit samples to estimate one complex scalar) and removes a 1–3 dB kink at $K \leq 4$ that the unboosted curriculum left behind on the N210 sweep. The synthetic Rician+IQ checkpoint uses p_{burst}=0. Each batch item draws an independent PA, SI channel, I/Q imbalance, and AWGN realization, so within a batch every sequence is a different ICL task.

A.2. Reverse Pseudo-Huber Loss

The reverse pseudo-Huber loss (15) of Odland (2024) is parametrized by a single scale $\delta > 0$. Writing $u = |e|/\delta$, its gradient is

$$\rho'_\delta(e) = \operatorname{sign}(e)\, \sqrt{u^2 + 1}, \tag{16}$$

which has constant magnitude as $|e| \to 0$ so $\rho'_\delta(0^\pm) = \pm 1$, an L_1-style sign step, and grows linearly with $|e|/\delta$ for large errors (an L_2-style linear ramp). This is the *reverse* of the standard (pseudo-)Huber, which is L_2 near the origin and L_1 in the tails. Figure 3 overlays ρ_δ with L_1 and L_2. The constant-magnitude gradient on small residuals is what we need at deep NMSE.

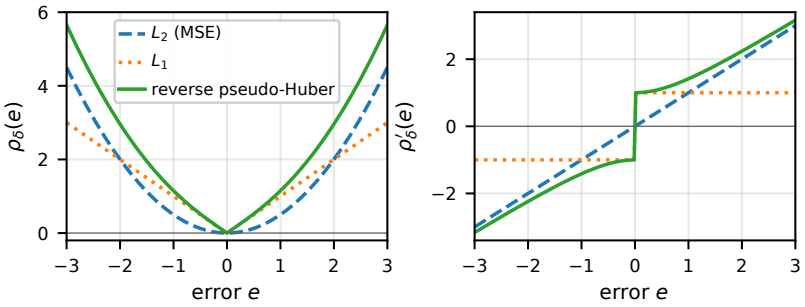


Figure 3. Loss (**left**) and gradient (**right**) of the reverse pseudo-Huber (δ=1) overlaid with L_1 and L_2. The reverse pseudo-Huber matches the constant-magnitude L_1 gradient near zero and transitions to the linear L_2 gradient in the tails.

B. Real PA with Synthetic Impairments

B.1. Reference PA Model

The reference coefficients $\boldsymbol{a}^\star$ are obtained by fitting the odd-order memory polynomial

$$y[n] \;=\; \sum_{p=0}^{P}\sum_{m=0}^{M} a^\star_{p,m}\, x[n-m]\,|x[n-m]|^{2p}$$

to the OpenDPD DPA-200 MHz training-split IQ pairs by ordinary least-squares (Morgan et al. 2006). To select $(P^\star, M^\star)$ we sweep $P\in\{1,3,5,7,9\}$ and $M\in\{0,\dots,5\}$ and evaluate held-out NMSE on the OpenDPD validation split. Validation NMSE saturates at $(P^\star, M^\star) = (9, 4)$ — 25 complex coefficients, NMSE -29.98 dB on validation and -30.04 dB on test.

B.2. Per-Sequence Task Draw (Figures 1a and 1b)

Figures 1a and 1b share a single checkpoint trained on the configuration in Table 2; Figure 1b only differs at evaluation, where complex-Gaussian queries are replaced by OFDM. Each sequence draws one coefficient perturbation $\boldsymbol{\varepsilon}$, shared across the K context samples and the query. Stacking the entries $a^\star_{p,m}$ row-major into $\boldsymbol{a}^\star\in\mathbb{C}^D$, the perturbed task coefficient is $\boldsymbol{a} = \boldsymbol{a}^\star + \boldsymbol{\varepsilon}$ with

$$\boldsymbol{\varepsilon}\sim\mathcal{CN}\big(\mathbf{0},\,\boldsymbol{\Sigma}_a\big),\quad \boldsymbol{\Sigma}_a \;=\; \sigma_a^2\,\mathrm{diag}\big(|a_1^\star|^2,\dots,|a_D^\star|^2\big).$$

Transmit samples are i.i.d. $x[n]\sim\mathcal{CN}(0,P_x)$ and AWGN $w[n]\sim\mathcal{CN}(0,\sigma_w^2)$ is added at the PA output.

Table 2. Per-sequence task-draw configuration for Figures 1a and 1b.

Reference $(P^\star, M^\star)$	$(9, 4)$, 25 complex coefficients
Coefficient drift σ_a	0.1
Transmit power P_x	0.1
Test waveform	Gaussian (Figure 1a) or QPSK-OFDM (Figure 1b)
SI channel	single-tap
I/Q imbalance	off
AWGN σ_w^2	10^{-2} (fixed; equiv. SNR $+10$ dB at P_x=0.1)

B.3. Real PA with Channel and I/Q (Figure 1c)

For Figure 1c the synthetic transmit-PA cascade is replaced with measured DPA-200 MHz IQ samples from the OpenDPD test split. Each sequence convolves the measured PA output with an independent Rician channel $\boldsymbol{h}\in\mathbb{C}^L$ (L=10, exponential PDP $\bar\sigma_\ell^2\propto e^{-0.5\ell}$ normalised to unit total power, K_{ric}=10 dB), passes the result through a widely-linear I/Q imbalance $y_{\mathrm{IQ}} = c_1 y + c_2 y^*$ with $c_1 = (1+ge^{j\varphi})/2$, $c_2 = (1-ge^{j\varphi})/2$, $|g_{\mathrm{dB}}|\sim\mathcal{U}(0.5, 3.0)$, $|\varphi|\sim\mathcal{U}(1^\circ, 5^\circ)$ (each with independent random sign), and adds AWGN at $\mathrm{SNR}_{\mathrm{dB}}\sim\mathcal{U}(-10, 20)$. ICL and the Bayesian WL-ALS baseline read the same windows of this pipeline.

C. Classical Baselines

RLS-MP. The classical baseline used in Figures 1a and 1b. It implements (9) verbatim: $\hat{\boldsymbol{a}} = \boldsymbol{a}^\star + (\boldsymbol{\Phi}^{\mathsf{H}}\boldsymbol{\Phi} + \lambda\boldsymbol{I})^{-1}\boldsymbol{\Phi}^{\mathsf{H}}(\boldsymbol{y} - \boldsymbol{\Phi}\boldsymbol{a}^\star)$. The estimator has access only to $\boldsymbol{a}^\star$ and the in-context calibration data; it does *not* see the perturbation prior $\boldsymbol{\Sigma}_a$,

so the ridge weight λ on $\|\boldsymbol{a} - \boldsymbol{a}^\star\|^2$ is a free hyperparameter. We sweep $\lambda \in \{10^{-2}, 5\cdot10^{-2}, 10^{-1}, 3\cdot10^{-1}, 1, 3, 10\}$ on a held-out task draw and select the value that minimizes NMSE at $K{=}K_{\max}$, which gives $\lambda{=}1$. This corresponds to the strongest scalar-ridge approximation to the prior, isotropic $\mathcal{CN}(\boldsymbol{a}^\star, \sigma_w^2\boldsymbol{I})$; it is still strictly weaker than the diagonal $\boldsymbol{\Sigma}_a = \sigma_a^2\,\mathrm{diag}(|\boldsymbol{a}^\star|^2)$ that the Bayes-optimal predictor exploits, since for the DPA reference $|\boldsymbol{a}_i^\star|^2$ varies by $\sim$600$\times$ across coefficients, which no scalar λ can mimic. The residual gap between RLS-MP and Bayes-optimal in Figures 1a and 1b is therefore the information advantage of knowing $\boldsymbol{\Sigma}_a$. For the hardware results, the per-context matched-filter gain $\hat{\alpha}$ from (14) is also applied to $\boldsymbol{a}^\star$ so that its units are commensurate with the per-device data.

Bayesian widely-linear ALS. The classical baseline used in Figure 1c. The PA branch is initialised at $\boldsymbol{a}^\star$ and the conjugate (image) branch at $\boldsymbol{a}^{\star *}$, with a Gaussian MAP anchor of variance $(\sigma_a|\boldsymbol{a}^\star|)^2$ matched to the same coefficient-perturbation scale σ_a that the transformer was pre-trained on. The widely-linear basis explicitly models both the direct ($\alpha\, y$) and the image ($\beta\, y^*$) components of the I/Q imbalance, so the estimator is structurally matched to the data-generating process. The channel is initialized at the identity tap and ridge-regularized with $\lambda{=}\sigma_w^2\,\hat{P}_y$, where $\hat{P}_y$ is the in-context received-signal power; this anchor breaks the $(\boldsymbol{a}\,c, \boldsymbol{h}/c)$ bilinear ambiguity that would otherwise destabilize the alternation at small K. Five alternating updates are sufficient for convergence.

D. Hardware Capture Details

The N210 captures are centered at 915 MHz with a 25 MS/s sample rate. Tx and Rx gain are pinned at $\mathrm{tx\_gain} = 31.5$ dB and $\mathrm{rx\_gain} = 0$ dB across all sessions and devices, so per-device variation in the receive path appears in the per-context gain $\hat{\alpha}$ rather than in a manually-tuned operating point. An external RF attenuator is inserted in the direct loopback path to keep receive samples below the ADC full-scale at the highest transmit scales we sweep.

Transmit scale σ_x and signal bandwidth fraction β are swept across a grid of values to induce operating-point variation. Each capture is segmented into non-overlapping windows containing K context examples and one query sample, with the per-window stride set to $d = \lceil 1/\beta \rceil$ as in Appendix 4 (Appendix G ablates this choice). All normalization statistics used for a query — including the per-context gain $\hat{\alpha}$ in (14) — are computed from its own context only.

E. N210 Power-Amplifier Characterization

A power amplifier (PA) is the last active stage of the transmit chain and is weakly nonlinear: at low drive its input–output map is memoryless and linear, but as drive approaches saturation the gain compresses and out-of-band intermodulation grows. These nonlinearities are the dominant source of residual SI once analog cancellation has removed the linear leakage, which is why Equation (8) models them as a memory polynomial.

Methods. We characterise the N210 PA on Device 1 in three complementary ways. *(i)* A continuous-wave (CW) tone at $f_c + 2$ MHz is swept over 40 logarithmically spaced tx scales (0.005 to 0.95) with tx_gain held fixed at 31.5 dB, and a single complex gain $\alpha = \boldsymbol{x}^{\mathsf{H}}\boldsymbol{y}/\boldsymbol{x}^{\mathsf{H}}\boldsymbol{x}$ is fit per drive level. *(ii)* For each $\sigma_x \in \{0.20, 0.30, \ldots, 0.95\}$ a 20 MHz band-limited complex Gaussian is transmitted continuously and the average P_{out} is read off the spectrum analyzer as channel power. *(iii)* A single 20 MHz Gaussian capture at $\sigma_x{=}0.7$ is aligned to fractional-sample precision and binned by instantaneous $|x|$; the per-bin mean $|y|$ yields a sample-level AM/AM curve. Output-side dBm is anchored to the small-signal CW point ($\sigma_x{=}0.03$, $P_{\mathrm{out}}{=}+2.83\,\mathrm{dBm}$). The three views differ in how σ_x loads the PA: CW has peak = rms = σ_x, while Gaussian has peak $\approx \sigma_x$ but rms $\approx \sigma_x/4.04$, so at the same numeric σ_x CW puts $\sim$12 dB more average power into the PA than Gaussian.

Results. Figure 4 shows the three curves. The CW curve (left) grows roughly 20 dB per decade up to $\sigma_x \approx 0.1$, crosses $P_{1\mathrm{dB}}$ near $\sigma_x{\approx}0.235$, and saturates at $P_{\mathrm{out}} \approx +20$ dBm by $\sigma_x{\approx}0.30$. The Gaussian average-power curve (middle) tracks the small-signal slope through $\sigma_x{=}0.60$ to within 0.5 dB and crosses $P_{1\mathrm{dB}}$ only at $\sigma_x{\approx}0.78$ ($P_{\mathrm{out}}{\approx}+17.4$ dBm). The instantaneous-$|x|$ panel (right) departs from its small-signal slope earlier than CW and saturates at a similar P_{out} — the expected peak-clipping signature of a $\sim$12 dB-PAPR envelope, with instantaneous Gaussian peaks crossing P_{sat} well before the RMS does.

Operating regime. The Gaussian average-power curve is the operative one for the experiment grid in Appendices H and I. Concretely, σ_x values up to 0.60 are at most $\sim$0.5 dB from the small-signal extrapolation, and only $\sigma_x{=}0.70$ and beyond

enter the immediate vicinity of $P_{1\mathrm{dB}}$.

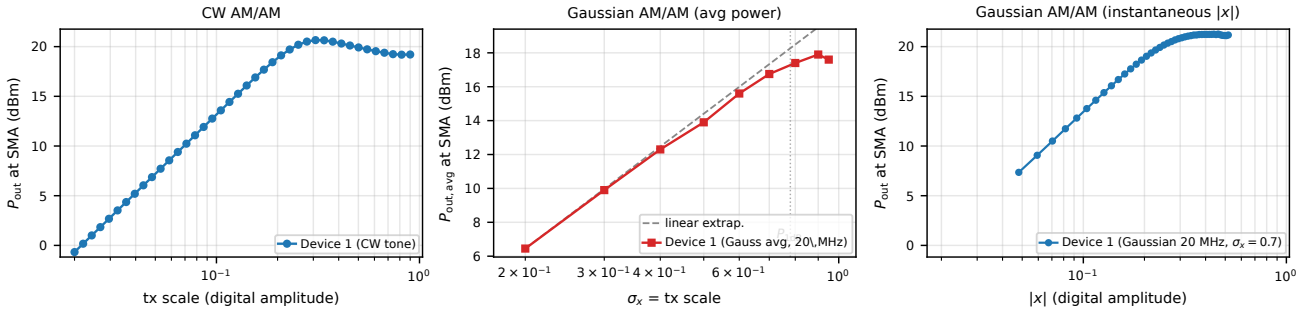


Figure 4. Device 1 PA characterization, with tx_gain held fixed at 31.5 dB. **Left:** CW AM/AM (single tone at $f_c + 2$ MHz), 40 logarithmically spaced tx scales; CW has peak = rms = tx scale, so the curve saturates by tx scale ≈0.30. **Middle:** Gaussian *average*-power AM/AM, $P_{\mathrm{out,avg}}$ at the SMA versus σ_x= tx scale (spectrum-analyzer channel-power readings of a 20 MHz band-limited complex Gaussian; PAPR ≈12 dB so rms ≈ tx scale/4.04). Dashed grey is the small-signal linear extrapolation from σ_x=0.20; $P_{1\mathrm{dB}}$ is at σ_x≈0.78. *This is the curve relevant to the σ_x labels used in the experiment grid: at σ_x=0.60 the average-power compression is only ∼0.4 dB.* **Right:** sample-level AM/AM from the same Gaussian capture (at σ_x=0.7), binned by instantaneous $|x|$. CW and instantaneous panels share a y-axis; the middle panel sits on a different drive scale (rms vs. peak) and uses its own range. Output-side dBm anchored to the small-signal CW measurement (σ_x=0.03, $P_{\mathrm{out}} = +2.83$ dBm).

F. SNR Ablation

Figure 5 sweeps the AWGN level on the Gaussian-perturbation, single-tap DPA-200 MHz benchmark. ICL tracks the analytic Bayes-optimal MSE within 1 dB for SNR ≤ 14 dB across $K \in \{16, 64, 128\}$ and saturates at high SNR.

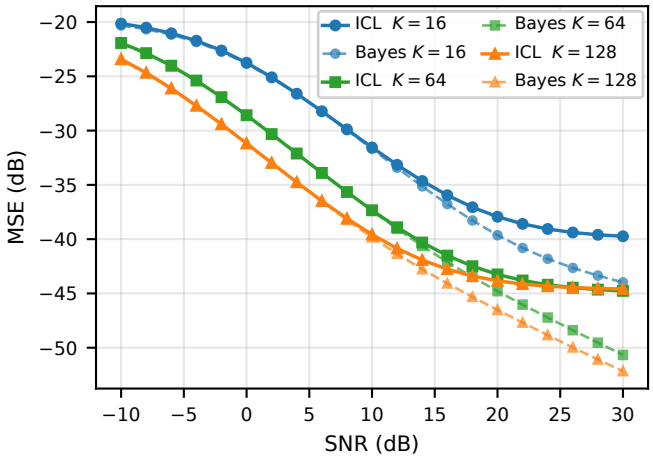


Figure 5. Synthetic SNR ablation, DPA-200 MHz, Gaussian coefficient drift. MSE versus AWGN SNR for $K \in \{16, 64, 128\}$. Solid: ICL. Dashed: analytic Bayes-optimal.

G. Effect of Context Decimation

Figure 6 examines a low-bandwidth, low-transmit-scale Device 1 cell ($B \approx 2.5$ MHz, σ_x=0.2), where the per-window stride $d = 10$ is largest and the consequences of correlated context are most visible. Without the stride, the in-context predictor and the linear RLS baselines first improve (up to $K \approx 4$) and then *degrade* with additional context, plateauing around $K \approx 16$. The cause is the same for both estimators. As the per-window stride shrinks below $1/\beta$, neighboring context windows $\boldsymbol{x}_i, \boldsymbol{x}_{i+1}$ become near-identical, so the empirical Gram matrix of the linear basis becomes ill-conditioned and the prior-warm-started RLS solution drifts back toward a copy of the most recent $(\boldsymbol{x}_i, y_i)$; the transformer head settles into the same shortcut, learning to copy the matched y of the closest context input rather than to identify the underlying PA. With the bandwidth-matched stride applied at training and inference, that redundancy disappears and ICL beats every classical baseline across $K \in [1, 128]$, improving monotonically down to ~ -35 dB.

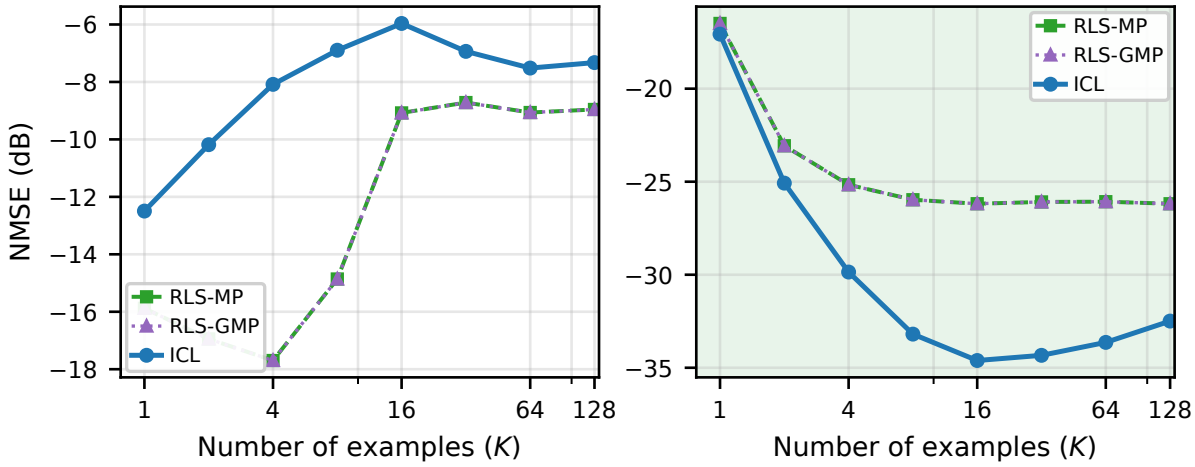


Figure 6. Effect of bandwidth-matched context decimation, low-bandwidth Device 1 cell ($B \approx 2.5$ MHz, σ_x=0.2). **Left:** same cell processed at the raw ADC rate (d=1). **Right:** bandwidth-matched stride $d = \lceil 1/\beta \rceil$ at training and inference (latest paper checkpoint). Lower is better.

H. Per-cell Hardware Curves: Same-device

Figure 7 expands Figure 2a along the transmit-scale axis at the same signal bandwidth (16.25 MHz), sweeping $\sigma_x \in \{0.40, 0.50, 0.70\}$ to cover the linear and soft-compression regimes (PA $P_{1\mathrm{dB}}$ sits at $\sigma_x \approx 0.78$, so all three cells are sub-$P_{1\mathrm{dB}}$). The ICL curve sits below the classical baseline in every cell, with the gap widening as σ_x pushes the PA into compression.

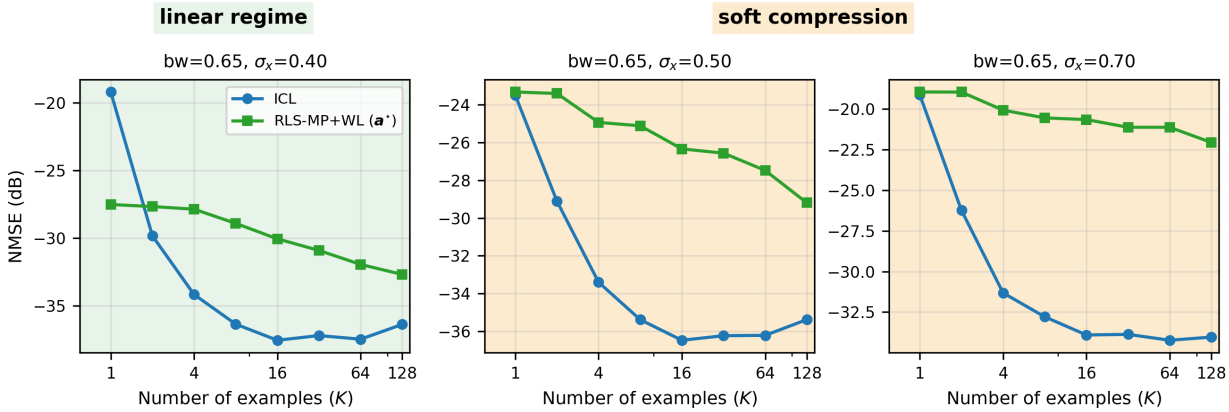


Figure 7. Device 1 per-cell NMSE versus K at signal bandwidth 16.25 MHz for three transmit scales. Background tint marks the PA operating regime: **linear** (σ_x=0.40, ∼0.2 dB compression, green) and **soft compression** ($\sigma_x \in \{0.50, 0.70\}$, ∼0.5–0.6 dB compression, orange); PA $P_{1\mathrm{dB}}$ at $\sigma_x \approx 0.78$. RLS-MP+WL is warm-started with the per-cell asymptotic $\boldsymbol{a}^\star$ obtained from a large-K (K=10^5) LSQ fit on the matching Device 1 training capture. Lower is better.

I. Per-cell Hardware Curves: Cross-device

Figures 8 and 9 report the per-cell breakdowns for the two cross-device test units corresponding to Figure 2b.

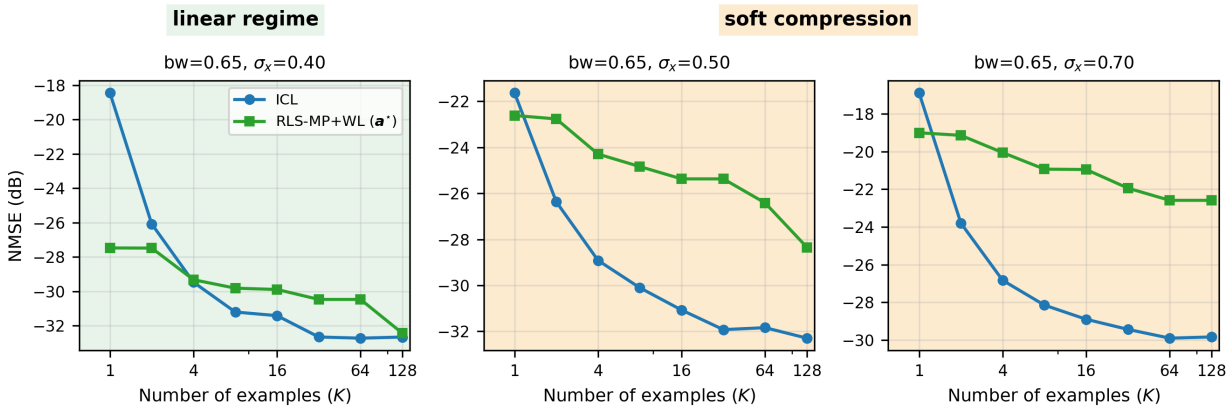


Figure 8. Device 2 per-cell NMSE versus K at signal bandwidth 16.25 MHz for $\sigma_x \in \{0.40, 0.50, 0.70\}$, zero-shot from the Device 1 ICL checkpoint and the Device 1 per-cell $\boldsymbol{a}^\star$ warm-start. Same regime tint convention as Figure 7. Lower is better.

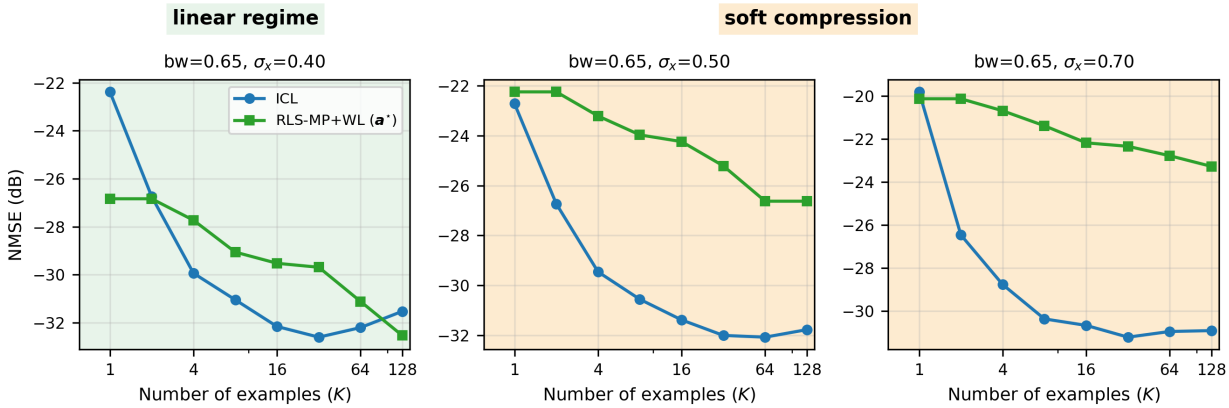


Figure 9. Device 3 per-cell NMSE versus K at signal bandwidth 16.25 MHz for $\sigma_x \in \{0.40, 0.50, 0.70\}$, zero-shot from the Device 1 ICL checkpoint and the Device 1 per-cell $\boldsymbol{a}^\star$ warm-start. Same regime tint convention as Figure 7. Lower is better.

J. Per-cell Hardware Curves: $\hat{\alpha}$ Ablation

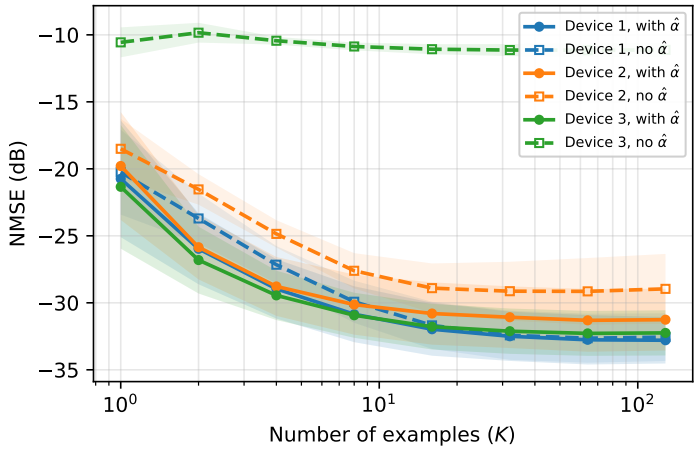


Figure 10. Per-context $\hat{\alpha}$ ablation summary, all three devices. With-$\hat{\alpha}$ production checkpoint vs. a from-scratch no-$\hat{\alpha}$ retrain on the same data. Lower is better.

The per-window gain $\hat{\alpha}$ from (14) varies substantially across units in both magnitude and phase, so a model that absorbs device gain into its weights cannot be expected to transfer. Figure 10 ablates the $\hat{\alpha}$ projection by retraining a no-$\hat{\alpha}$ checkpoint from scratch on the same data. On the training device the two checkpoints agree closely; on a second unit a small gap opens up; on the third unit, where the per-session $\hat{\alpha}$ phase rotates substantially relative to the training device, the no-$\hat{\alpha}$

NMSE flattens across the entire K sweep, indicating that the model is no longer performing in-context learning at all. The $\hat{\alpha}$ projection is therefore not a numerical convenience: it factors out the per-device linear variability that does not transfer across hardware units, leaving the transformer free to model the nonlinear PA structure that does.

Figures 11 to 13 expand Figure 10 into the full 5×5 grid for each device, overlaying the with-$\hat{\alpha}$ production checkpoint and the from-scratch no-$\hat{\alpha}$ retrain. The qualitative picture in the body figure holds at the cell level. On Device 1 the two checkpoints overlap closely across the grid, with the no-$\hat{\alpha}$ curve retaining the same bend in K. On Device 2 the no-$\hat{\alpha}$ retrain trails by a few dB but tracks the with-$\hat{\alpha}$ shape. On Device 3 the no-$\hat{\alpha}$ curves are essentially flat in K in every cell of the grid, sitting near -11 dB regardless of bandwidth or transmit scale; the with-$\hat{\alpha}$ checkpoint, in contrast, retains the same K-dependent bend it shows on Devices 1 and 2.

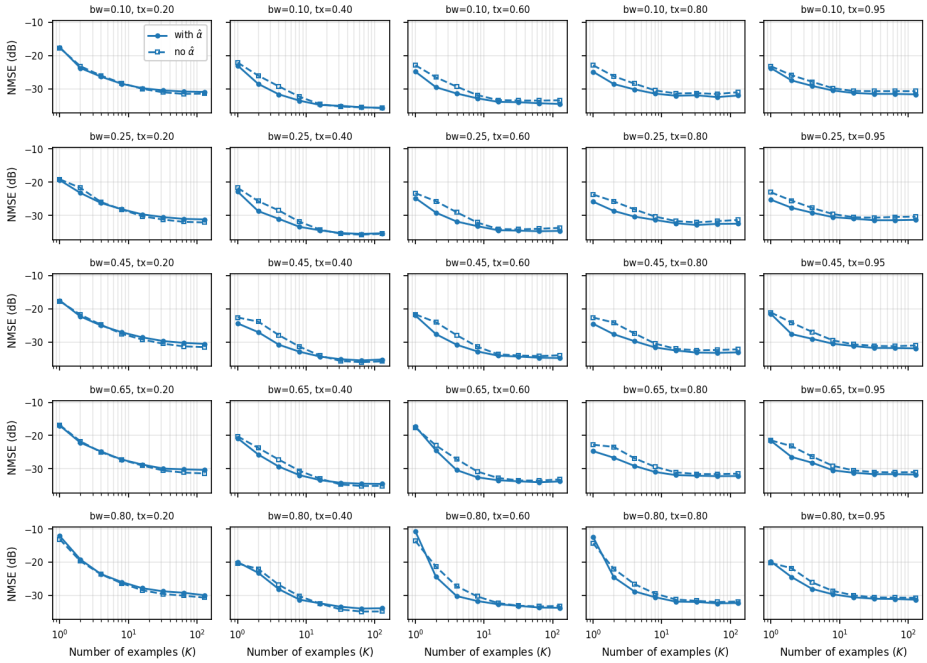


Figure 11. Device 1, per-cell NMSE versus K for the with-$\hat{\alpha}$ checkpoint and the no-$\hat{\alpha}$ retrain. Lower is better.

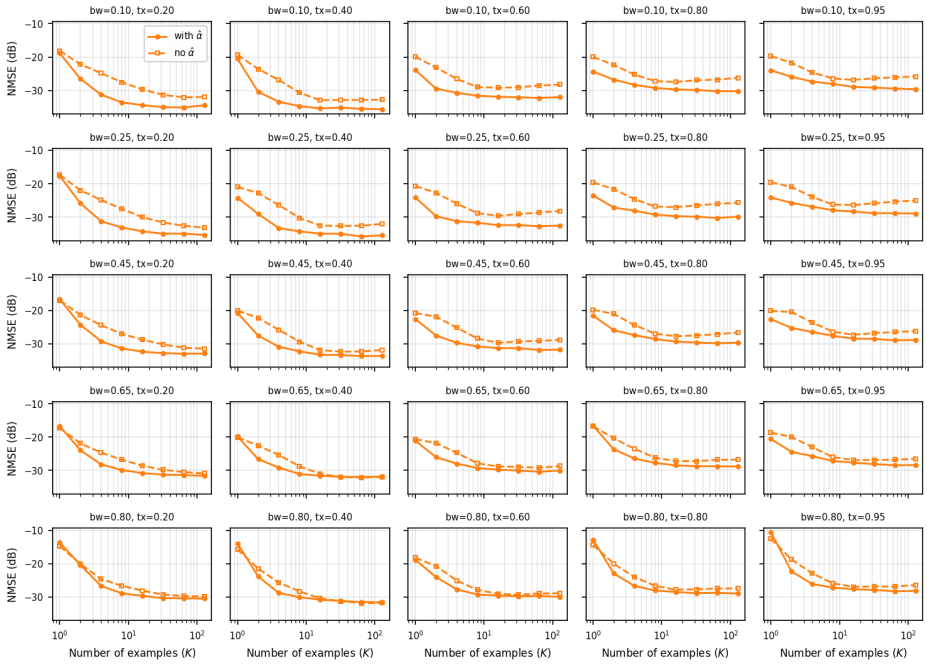


Figure 12. Device 2, per-cell NMSE versus K for the with-$\hat{\alpha}$ checkpoint and the no-$\hat{\alpha}$ retrain. Lower is better.

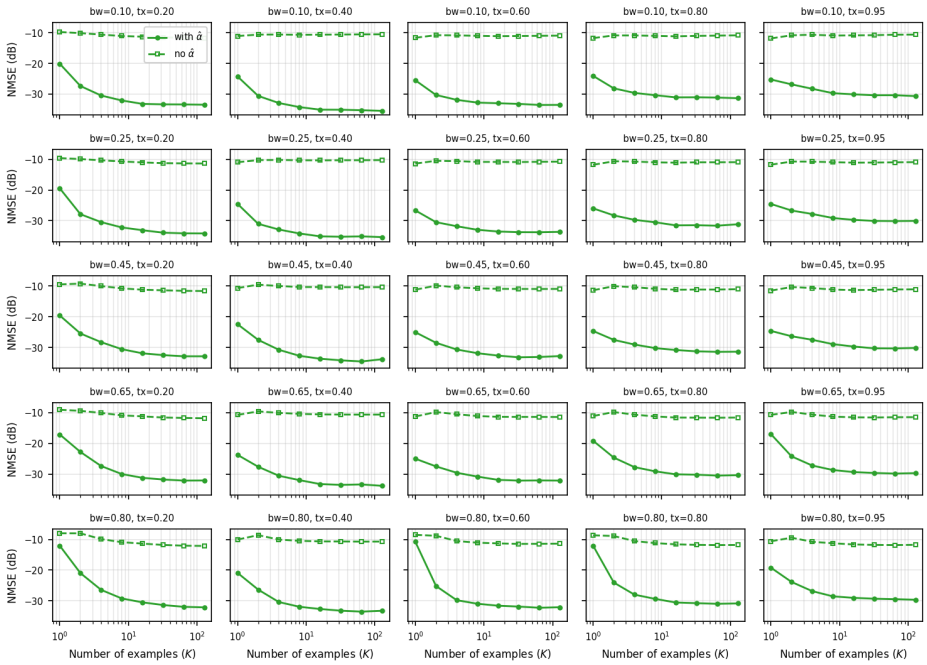


Figure 13. Device 3, per-cell NMSE versus K for the with-$\hat{\alpha}$ checkpoint and the no-$\hat{\alpha}$ retrain. Lower is better.