# Multi-Task Transformer Receiver for OFDM Channel Estimation and Symbol Detection

**Zhoubin Kou, Renpu Liu, Jing Yang, Cong Shen**
University of Virginia
Charlottesville, VA 22903
{zhoubin, pzw7bx, yangjing, cong}@virginia.edu

## Abstract

By leveraging in-context learning (ICL), pretrained Transformers adapt to unseen tasks from example prompts without task-specific fine-tuning. This adaptability has motivated their use in wireless communications, where ICL-based Transformers have shown strong performance on symbol detection. However, deploying a Transformer solely for symbol detection is less cost-effective. Can we design a multi-task Transformer that, without significantly increasing inference overhead, unifies additional modules of the wireless communication receiver within a single model? In this work, we propose a multi-task ICL Transformer that treats pilots within a coherence block as prompts, and jointly outputs the detected data symbol and an explicit channel-frequency response (CFR). Empirically, we find that activating the model's multi-task capability improves both training efficiency and receiver performance under the same model size, compared to an ICL-based Transformer performing symbol detection alone.

## 1 Introduction

Transformers have recently been shown to perform *in-context learning* (ICL) [1, 2]: from a short prompt of input–output pairs, a pre-trained model can infer task structure and generalize to a held-out query without parameter updates. While most existing demonstrations originate from language and vision, emerging evidence indicates that ICL is also effective for physical-layer reception at a wireless receiver, where symbol detection can be posed as learning a mapping from pilots to symbol equalization during the online inference process [3–5]. This perspective is appealing in wireless settings with heterogeneous channels and scarce priors, because a single amortized model can rapidly adapt *in context* to the current realization rather than relying on explicit per-scenario retraining.

However, two practical gaps remain. First, pure implicit equalization[3–5]—where the model directly outputs detected symbols—can be both data-intensive and computationally demanding; adding structure that reflects the underlying channel often improves training efficiency and stability. Second, many downstream tasks in modern wireless systems still require an *explicit* channel estimate (e.g., channel-quality indicators (CQI) for feedback and integrated sensing and communications (ISAC) [6]). Methods that only perform implicit symbol detection limit interoperability and downstream utility. These considerations motivate a formulation in which a single model both exploits ICL to adapt from pilots in the inference process and exposes an explicit channel-frequency response (CFR) alongside symbol decisions.

We therefore study a multi-task, decoder-only Transformer that treats pilots within a coherence block as a $k$-shot support set and jointly outputs the detected data symbol and an explicit CFR for the query. Adding explicit channel estimation to an ICL model is both natural and principled. In classical receivers, channel estimation serves as the intermediate step for demodulation, which means the two tasks are not strictly parallel but instead form an estimate-then-detect progression. Treating channel

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: AI4NextG.

estimation (CE) as an auxiliary or intermediate task therefore aligns with the classical structure and provides a structured internal representation that holds promise for improving training efficiency and stability.

We focus on an orthogonal frequency-division multiplexing (OFDM) system because its standardized pilot structures furnish a high-resolution CFR and a clean interface to downstream modules (e.g., equalization, CQI, and sensing), while enabling controlled $k$-shot protocols that isolate genuine in-context adaptation from mere sequence-length effects. At inference time, we construct the input to the Transformer model by concatenating $k$ pilot pairs and a query instance. From the output sequence, the model simultaneously predicts both the transmitted symbol and the corresponding channel estimate. This unified output allows either end-to-end detection or classical equalization to be applied downstream. Our contributions are summarized as follows.

- We formulate joint channel estimation and symbol detection as one ICL problem in a wideband OFDM system, and design a multi-task Transformer that outputs both the detected symbol and an explicit CFR, furnishing a clear interface for downstream tasks that rely on channel state information.
- We introduce an explicit channel estimation task into the ICL formulation, which empirically accelerates convergence of the Transformer during training; concurrently, we train in a $k$-shot manner (sampling pilots of varying lengths as training examples) to improve robustness across context lengths and prevent overfitting to any single pilot length.
- Experimental results indicate that the multi-task ICL model achieves near-ideal performance on both symbol detection and channel estimation. Interestingly, it outperforms same-size ICL models trained only for symbol detection. Adding channel estimation enables an estimate-then-detect workflow, speeds up convergence with only a small inference cost, and improves online inference and downstream applicability.

## 2 System Model and Problem Formulation

### 2.1 Wireless Model

We consider the single-input single-output (SISO) OFDM system consisting of $N$ orthogonal subcarriers operating in a frequency-selective fading environment. Under the assumption of block fading [7], the CFR is considered invariant over a coherence interval spanning $B$ consecutive OFDM symbols, changing independently between intervals. Within the $b$-th coherence interval, the frequency-domain received signal vector at the $t$-th OFDM symbol, denoted as $\boldsymbol{y}^t \in \mathbb{C}^N$, can be expressed as:

$$\boldsymbol{y}^t = \boldsymbol{D}_H^b \, \boldsymbol{x}^t + \boldsymbol{n}^t, \quad t = 1, \ldots, B,$$

where $\boldsymbol{x}^t = [x_0^t, x_1^t, \ldots, x_{N-1}^t]^T \in \mathbb{C}^N$ represents the transmitted frequency-domain symbol vector drawn from a predefined constellation $\mathcal{X}$ (e.g., QPSK or 16-QAM), and $\boldsymbol{n}^t \sim \mathcal{CN}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_N)$ is a noise vector of independent and identically distributed (i.i.d.) additive white Gaussian noise (AWGN) samples with variance $\sigma^2$. The diagonal matrix $\boldsymbol{D}_H^b = \mathrm{diag}\{\boldsymbol{H}_b\}$ ($\boldsymbol{H}_b = [H_0^b, H_1^b, \ldots, H_{N-1}^b]$) characterizes the channel's frequency response at each of the $N$ subcarriers for the $b$-th coherence block. Due to the block-fading assumption, $\boldsymbol{D}_H^b$ remains constant over all OFDM symbols within the coherence interval but varies independently from one block to another.

### 2.2 Classical CE&SD Pipeline

Channel estimation is typically conducted using pilot symbols transmitted on a subset of subcarriers or within dedicated OFDM symbols. Specifically, if $N_p$ pilot symbol vectors $\{\boldsymbol{x}_p^i\}i = 1^{N_p}$ are transmitted within each coherence interval, the corresponding received pilot signals $\{\boldsymbol{y}_p^i\}i = 1^{N_p}$ can be collected to estimate the frequency response. An element-wise least-squares (LS) estimate of the channel coefficients $H_k^b$ is then obtained as:

$$\hat{H}_k^b = \frac{\sum_{i=1}^{N_p} y_{p,k}^i x_{p,k}^{i*}}{\sum_{i=1}^{N_p} |x_{p,k}^i|^2}, \quad k = 0, \ldots, N-1,$$

forming the diagonal estimate matrix $\hat{\boldsymbol{D}}_H^b = \mathrm{diag}\{\hat{H}_0^b, \ldots, \hat{H}_{N-1}^b\}$.

Given this channel estimate, symbol detection for data-bearing OFDM symbols is accomplished by applying a linear equalizer based on the inverse of the estimated channel. The equalized symbol estimates for the received vector $\boldsymbol{y}_d^t$ (excluding pilots) within the same coherence interval are computed as:

$$\tilde{\boldsymbol{x}}^t = \left(\hat{\boldsymbol{D}}_H^{b-1}\right)^{-1} \boldsymbol{y}_d^t,$$

followed by a constellation-based decision step:

$$\hat{\boldsymbol{x}}_k^t = \arg\min_{s \in \mathcal{S}} |\tilde{\boldsymbol{x}}_k^t - s|^2, \quad k = 0, \ldots, N-1,$$

where $\mathcal{S}$ represents the modulation symbol set. This frequency-domain formulation clearly delineates the pilot-based channel estimation and symbol-detection stages in a SISO OFDM system under block-fading conditions.

In contrast to the classical CE→EQ→SD pipeline in conventional wireless receivers, which first estimates the channel, then equalizes the received signal, and finally performs symbol detection, our approach integrates these steps into a single Transformer-based model that jointly produces channel estimates and detected symbols within one unified inference pass.

# 3 In-Context Learning-Based Joint Channel Estimation and Symbol Detection

## 3.1 Problem formulation

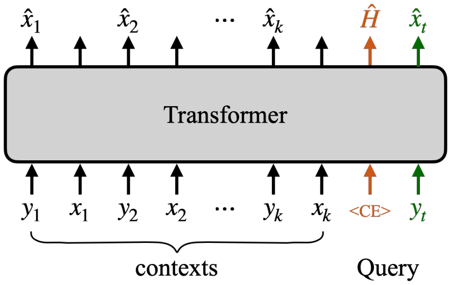

Figure 1: Inference prompt for our proposed decoder-only Transformer.

In this paper, we propose a multi-task ICL Transformer receiver for joint channel estimation and symbol detection in SISO OFDM systems. We leverage the Transformer's ability to learn from contexts (pilot information), allowing it to adaptively estimate the channel and detect symbols without requiring parameter updates at inference stage.

We cast joint channel estimation and symbol detection within a single coherence block $b$ as an in-context episode. As illustrated in Fig. 1, the receiver observes $N_p = k$ pilot pairs

$$\mathcal{D}^b = \left\{(\boldsymbol{x}_i^b, \boldsymbol{y}_i^b)\right\}_{i=1}^k, \qquad \boldsymbol{y}_i^b = \boldsymbol{D}_H^b\, \boldsymbol{x}_i^b + \boldsymbol{n}_i^b,$$

Together with a query that contains a special token $\langle\text{CE}\rangle$ indicating channel estimation followed by the received signal $\boldsymbol{y}_t$ to be detected. Given a Transformer model parameterized by $\boldsymbol{\theta}$, our goal is to predict both the channel matrix $\hat{\boldsymbol{H}}_b$ and detected symbol $\hat{\boldsymbol{x}}_t^b$. First, we design the input sequence $\boldsymbol{S}_{\text{input}}$ of the model as

$$\boldsymbol{S}_{\text{input}} := [\boldsymbol{y}_1^b, \boldsymbol{x}_1^b, \ldots, \boldsymbol{y}_k^b, \boldsymbol{x}_k^b, \langle\text{CE}\rangle, \boldsymbol{y}_t^b, ] \in \mathbb{R}^{(2k+1)\times 2N},$$

where each token concatenates the real and imaginary parts per subcarrier (except the $\langle\text{CE}\rangle$ token). Given the input sequence, the corresponding output sequence of the transformer model $f_\theta$ is $f_\theta(\boldsymbol{S}_{\text{input}})$. Ass shown in Fig. 1, we could obtain $\{\hat{\boldsymbol{x}}_i^b\}_{i=1}^k, \hat{\boldsymbol{H}}_b, \hat{\boldsymbol{x}}_t^b$ from the output sequence.

We employ a weighted multi-task loss to jointly optimize symbol detection and channel estimation. Specifically, the overall objective consists of three components: a context loss for the in-context exemplars, a detection loss for the target symbol, and a channel estimation loss.

$$\mathcal{L} = w_{\text{ctx}}\, \mathcal{L}_{\text{sd}}\left(\hat{\boldsymbol{x}}_i^b, \boldsymbol{x}_i^b\right) + w_{\text{sd}}\, \mathcal{L}_{\text{sd}}\left(\hat{\boldsymbol{x}}_t^b, \boldsymbol{x}_t^b\right) + w_{\text{ce}}\, \mathcal{L}_{\text{ce}}\left(\hat{\boldsymbol{H}}_b, \boldsymbol{H}_b\right),$$

where $\mathcal{L}_{\text{sd}}$ is the mean squared error (MSE) loss applied to both the context exemplars and the target symbol to ensure accurate symbol recovery, and $\mathcal{L}_{\text{ce}} = \|\hat{\boldsymbol{H}} - \boldsymbol{H}\|_2^2 / \|\boldsymbol{H}\|_2^2$ is the normalized mean squared error (NMSE) for channel estimation, which scales the error relative to the channel power so that large-magnitude coefficients do not dominate the training. The weights $w_{\text{ctx}}, w_{\text{sd}}, w_{\text{ce}}$ control the relative importance of these terms, allowing the model to balance detection and estimation performance without biasing toward a single task.

### 3.2 Multi-task ICL-based Transformer Model

**Model configuration**. We adopt a decoder-only Transformer backbone built on a GPT-2–style implementation (see Fig. 2). The backbone processes a single, serialized token sequence (interleaved pilot observation / pilot symbol pairs, a dedicated control token $\langle\text{CE}\rangle$, and optional test/query token $\boldsymbol{y}_t$) and returns the last-layer hidden states for every input token in one forward pass. Inputs are real-/imag-concatenated per subcarrier and projected into embedding space by a learned read-in layer; symmetric read-out layers linearly map selected hidden states back to the real-stacked complex representation used for losses and evaluation.

Causal masking and a prompt-aware (per-sample) attention mask are supplied to the backbone so that padded positions do not participate in attention and information flow remains directional. We deterministically lay out the sequence as interleaved pilot observation/symbol pairs followed by the $\langle\text{CE}\rangle$ control token and optional test token(s), and pad pilot pairs to an internal upper bound $K_{\max}$ so any $k \leq K_{\max}$ can be provided without changing model shape. In this layout the $\langle\text{CE}\rangle$ index is fixed across the batch: the $\langle\text{CE}\rangle$ hidden state may attend only to preceding pilot tokens (forming a pilot-conditioned representation), while the test/query token $\boldsymbol{y}_t$ may attend to both the pilot region and the $\langle\text{CE}\rangle$ state. Task heads are simple linear readouts on the backbone's last-layer vectors: the SD head maps each symbol-query hidden state to a complex symbol prediction (MSE training / inference) and the CE head projects the $\langle\text{CE}\rangle$ hidden state to a $2N$-dimensional real vector representing the CFR (used for NMSE and downstream CSI tasks). The implementation is fully vectorized so pilot, $\langle\text{CE}\rangle$ and test predictions are computed in batch for throughput and numerical stability. Overall, this design enforces an explicit estimate-then-detect inductive bias while preserving a single unified forward pass and efficient batching.

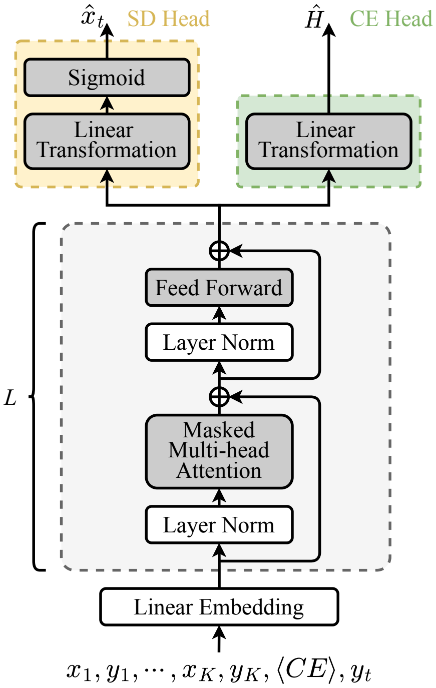

Figure 2: Decoder-only Transformer for ICL-based joint channel estimation and symbol detection.

**k-Shot Variable-Pilot Training**. Training is driven by an on-the-fly SISO-OFDM data generator that samples channel instances from a pre-constructed pool $\{(\boldsymbol{H}^b_\alpha, \sigma^2_\beta)\}, \alpha, \beta = 1, \cdots, n_{\text{task}}$. At the beginning of each epoch (and continuously during training), mini-batches are drawn by first selecting channel indices from this $n_{\text{task}}$ pool and then synthesizing the corresponding transmit symbols and noisy receive pilots under block-fading assumptions. For each generated example, the per-sample SNR is randomly drawn from a uniform range $[\text{SNR}_{\min},\ \text{SNR}_{\max}]$, so that the training set exposes the model to a diversity of noise conditions. To emulate realistic prompt variability, each sample's prompt contains a variable number of pilot pairs: during batching we randomly choose a pilot count $k$ from the interval $[1, K_{\max}]$ for each example, then collate-and-pad the resulting pilot pairs to the fixed tensor shape expected by the model. Padding is paired with a per-sample attention mask so that invalid pilot slots are ignored during both attention computation and loss accumulation. This procedure ensures that every training batch contains examples with heterogeneous pilot lengths and SNRs while keeping the forward pass fully vectorized and compatible with the deterministic $\langle\text{CE}\rangle$ token index used by the multi-task readouts.

## 4 Simulation Results

### 4.1 Settings

**Experimental setup.** We construct a channel pool of $n_{\text{task}} = 4096$ with a held-out test set of 1,000 examples. Training spans 15,000 epochs with 20 steps each (300,000 updates in total) using mini-batches of size 128. For each synthesized example, the channel input $s$ is drawn from a 4-QAM constellation $\mathcal{S}$, the per-sample SNR is uniformly sampled from $[\text{SNR}_{\min}, \text{SNR}_{\max}]$, and the number of pilot pairs $k$ is drawn from $[1, K_{\max}]$ with $K_{\max} = 30$. Pilot pairs are collated and padded to a fixed batch shape. The received signal is quantized by a mid-rise uniform quantizer with $b = 4$ bits and clipping range $[-4, 4]$. The OFDM system employs 64 subcarriers, whereas [8] considered at most 4. A linear warmup–decay learning-rate schedule is used, with warmup length $\max(1000, 0.01 \times 300{,}000) = 3{,}000$ steps.

**Model and optimization.** The backbone is a GPT-2–style decoder-only Transformer with embedding dimension 384, 8 layers, and 8 attention heads; the token feature dimension is data_dim $= 128$, context length is 30 and test length is 1. Optimization uses AdamW with $\mathrm{lr} = 1 \times 10^{-3}$, $\mathrm{weight\_decay} = 0.01$, $\beta = (0.9, 0.98)$ and $\epsilon = 1 \times 10^{-8}$; gradients are clipped to a maximum norm of 1.0. The multi-task loss weights are set as $w_{\mathrm{ctx}} = 0.5$, $w_{\mathrm{sd}} = 1.5$, and $w_{\mathrm{ce}} = 1.0$.

**Baselines.** We evaluate our multi-task ICL receiver against a set of classical and learning-based baselines under a matched pilot budget, identical sequence lengths, and with ICL-model weights frozen at test stage. Classical baselines comprise **LS+MMSE**, i.e., per-subcarrier least-squares estimation followed by MMSE equalization, and **LMMSE+MMSE**, a Bayesian variant that applies an LMMSE CFR estimator—with the prior covariance $\boldsymbol{R}_H$ tuned on validation—followed by MMSE equalization. As an upper reference we include **Oracle-MMSE**, which substitutes the true CFR into the MMSE equalizer. For learning baselines, we compare **SD-ICL (no CE head)**, a decoder-only Transformer that directly predicts the query symbol from the serialized pilot/query sequence without producing an explicit CFR. Our proposed **CE+SD (Ours)** multi-task ICL model jointly outputs in a single forward pass and can optionally feed the predicted CFR back into the SD path via pre-equalization or conditioning. Both ICL-based methods use the same tokenization and masking, and we match pilot ratios, SNR sampling ranges, and sequence lengths across experiments to ensure fair, apples-to-apples comparisons.

### 4.2 Training behavior analysis

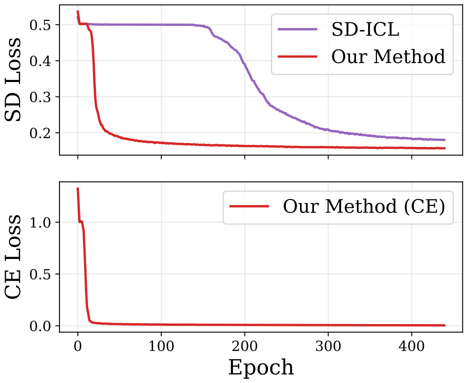

Figure 3: Test loss progress over training iterations.

As shown in Fig. 3, we compare the training dynamics of our method against the SD-ICL baseline. Across matched hyper-parameters and data, the proposed multi-task ICL receiver exhibits a clear two-stage convergence pattern. The channel-estimation test loss collapses in early epochs and plateaus near zero; only thereafter does the symbol-detection loss begin to decrease markedly. This ordering is consistent across seeds and SNR settings. These dynamics suggest that, for a Transformer trained via in-context learning, the pilot-to-CFR mapping is easier to induce than the pilot-to-decision mapping, and that supervising an explicit CFR yields better-conditioned representations for downstream detection. Relative to the SD-ICL baseline (no CE head), our method triggers the drop in SD loss tens to hundreds of epochs earlier and maintains a uniformly lower trajectory thereafter. We attribute the acceleration to the CE head, which injects an "estimate-then-detect" inductive bias and supplies early, low-variance gradients that organize hidden states around channel structure. Once this channel representation is formed, the SD head requires only modest additional optimization. Conceptually, the dynamics mirror step-by-step reasoning: the network first infers the intermediate variable (the channel) and then refines the final decision, akin to a chain-of-thought curriculum.

### 4.3 K-shot scaling analysis

Figure 4 (left) plots channel-estimation NMSE (dB) and (right) the corresponding BER as a function of the number of pilot pairs $k$ used by each method. From the $k$-shot experiments, we observe that as the number of pilots (contexts) increases, our models exhibit steady performance improvements. This validates that the trained SD-ICL and multi-task ICL receivers indeed activate their in-context learning capability rather than merely memorizing task-specific mappings for symbol detection or channel estimation. Such behavior ensures that the ICL-based models can be directly deployed under varying channel conditions without retraining.

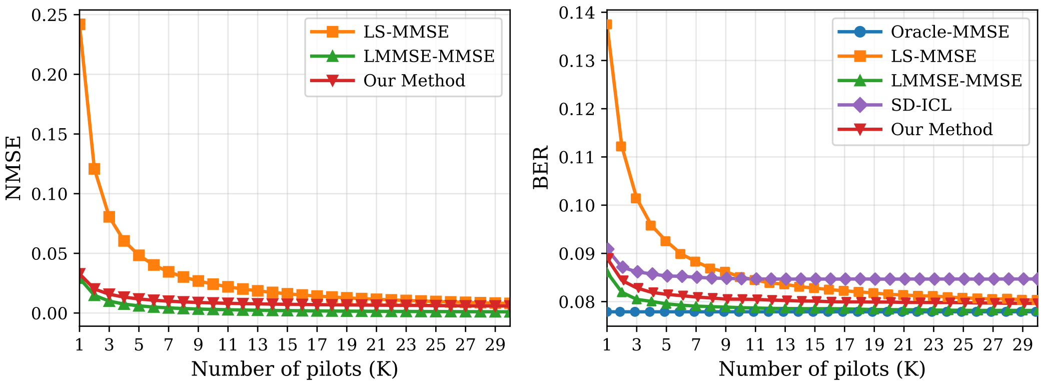

Figure 4: Impact of pilot length $k$: (Left) channel estimation, (Right) symbol detection.

Our multi-task CE+SD ICL model yields substantially lower MMSE across the full range of $k$ compared to the classical baselines LS: the gap is largest at very small $k$ (pilot-constrained regime) and narrows as $k$ increases, indicating diminishing returns from additional pilots once a modest pilot budget is available. By contrast, the classical LS and LMMSE estimators track each other closely and improve only slowly with $k$. This behaviour shows that the multi-task model extracts more informative, low-variance channel representations from few pilots, which translates into markedly better NMSE in the pilot-scarce regime.

The BER curves corroborate the NMSE results and illuminate the role of explicit CFR prediction for detection. The Oracle-MMSE curve (ideal upper bound) remains lowest as expected. Our method consistently outperforms the SD-ICL baseline (which does not produce an explicit CFR) and the classical baselines for nearly all $k$, and it moves closer to the Oracle as $k$ increases. The SD-ICL baseline exhibits a relatively flat BER with increasing $k$, suggesting limited ability to capitalize on extra pilots when no explicit channel estimate is produced. Taken together, these results indicate that jointly training for CE and SD (i) accelerates and stabilizes learning by providing early, low-variance gradients from the CE head, (ii) produces CFR estimates that can be fed back (via pre-equalization or conditioning) to improve detection, and (iii) yields superior sample efficiency under tight pilot budgets.

## 5 Conclusion and Future Works

We formulated joint channel estimation and symbol detection as an ICL episode and instantiated a multi-task, decoder-only Transformer that consumes pilots within a coherence block as a support set and – in a single, zero-gradient forward pass – produces both a symbol decision and an explicit frequency-domain channel response. A prompt-aware, causal masking scheme and per-sample (pilot-length–aware) padding/masking enable fixed-shape batching while preserving an explicit estimate-then-detect inductive bias. Under matched model size and pilot budgets in the SISO-OFDM setting, the multi-task objective improves training efficiency and yields promising, baseline-competitive performance on both CE (NMSE) and SD (BER/SER), with particularly strong gains in the pilot-scarce regime. We also observe a consistent two-stage training dynamic in which CE converges early and stabilizes the representations used by SD. Collectively, these results indicate that exposing an explicit CFR alongside symbol decisions is a practical way to unify receiver functionality without materially increasing inference overhead in ICL-driven receivers.

Building on this work, several directions merit further exploration. First, a theoretical analysis [9] of how the CE head assists SD in multi-task ICL would clarify the source of performance gains, which in turn can guide the extension of ICL-based Transformers to broader multi-task capabilities and provide a theoretical foundation for their application and efficient deployment in wireless communication tasks. Second, task scope can be extended beyond CE and SD by using the CFR as a unifying interface for sensing and link adaptation, aiming for a single ICL backbone that serves multiple receiver functions without bespoke add-ons. Third, chain-of-thought–style formulations that render intermediate reasoning explicit present a path to lower capacity requirements for strong generalization and to improved controllability [10].

## Acknowledgments and Disclosure of Funding

The work of Zhoubin Kou and Cong Shen were partially supported by the U.S. National Science Foundation (NSF) under grants 2132700 and 2143559. The work of Renpu Liu and Jing Yang were partially supported by the U.S. National Science Foundation (NSF) under grant 2531789.

## References

[1] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[3] Li Fan, Jing Yang, and Cong Shen. Decision feedback in-context symbol detection over block-fading channels. *arXiv preprint arXiv:2411.07600*, 2024.

[4] Matteo Zecchin, Tomer Raviv, Dileep Kalathil, Krishna Narayanan, Nir Shlezinger, and Osvaldo Simeone. In-context learning for gradient-free receiver adaptation: Principles, applications, and theory. *arXiv preprint arXiv:2506.15176*, 2025.

[5] Vishnu Teja Kunde, Vicram Rajagopalan, Chandra Shekhara Kaushik Valmeekam, Krishna Narayanan, Jean-Francois Chamberland, Dileep Kalathil, and Srinivas Shakkottai. Transformers are provably optimal in-context estimators for wireless communications. In *International Conference on Artificial Intelligence and Statistics*, pages 1531–1539. PMLR, 2025.

[6] Fan Liu, Yuanhao Cui, Christos Masouros, Jie Xu, Tony Xiao Han, Yonina C Eldar, and Stefano Buzzi. 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.

[7] David Tse and Pramod Viswanath. *Fundamentals of wireless communication*. Cambridge university press, 2005.

[8] Zihang Song, Yuan Ma, Chaoqun You, Haoxuan Yuan, Jinbo Peng, and Yue Gao. Transformer-based adaptive ofdm mimo equalization in intelligence-native ran. In *2024 IEEE/CIC International Conference on Communications in China (ICCC Workshops)*, pages 179–184. IEEE, 2024.

[9] Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*, 2021.

[10] Li Fan, Peng Wang, Jing Yang, and Cong Shen. Chain-of-thought enhanced shallow transformers for wireless symbol detection. *arXiv preprint arXiv:2506.21093*, 2025.