LEMMO: A LARGE ELECTROMAGNETIC MODEL LEVERAGING PRETRAINED LLMS FOR MILLION-SAMPLE SIGNAL UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models have expanded beyond text to vision, audio, and other modalities, but understanding raw electromagnetic (EM) signals remains challenging. Wideband recordings can contain millions of samples, whereas existing waveform-based EM models are primarily evaluated on short in-phase and quadrature (IQ) segments, leaving long-recording understanding and detailed description insufficiently explored. We introduce LEMMO, a large EM model for instruction-driven recognition and description of long raw IQ recordings. LEMMO combines a linear-attention EM encoder with anti-aliased temporal resampling to process up to eight million complex samples. Its EM–text interaction module conditions learnable queries on instructions before retrieving signal evidence. Compact queries support recognition, while structured queries with auxiliary factual supervision organize scene-level properties, signal attributes, and cross-signal relationships for detailed description. We also construct and release a unified 1.38-TB corpus of real-world and synthetic EM signals, including LEMMO-ShortQA and LEMMO-LongDesc. Compared with MERLIN-LoRA, LEMMO improves mean EM-Bench perception accuracy by 15.77 %, mean strategy-generation ROUGE-L F1 by 4.38 points, and mean LEMMO-ShortQA accuracy by 14.31 %. It also improves categorical and numerical fact F1 on both LEMMO-LongDesc tasks while preserving its language backbone's MMLU performance.

1 INTRODUCTION

Electromagnetic (EM) signals carry information about the wireless activities around us: which devices are transmitting, how they use the spectrum, and how their transmissions overlap or interfere. Interpreting this information supports applications across wireless communications, environmental sensing, autonomous systems, and security (Zhang et al., 2024; Dong et al., 2021; Yu et al., 2024b). Such understanding extends beyond isolated waveform classification to identifying signal components, characterizing their behavior, and interpreting their temporal and spectral relationships. Large language models (LLMs) have expanded beyond language to modalities such as vision, enabling natural-language interaction with increasingly diverse types of information Dai et al. (2023); Huang et al. (2023); Chen et al. (2024). Extending this capability to EM signals requires a corresponding interface that extracts instruction-relevant evidence from raw waveforms and incorporates it into response generation.

Existing research provides important steps toward this objective. Task-specific deep learning methods have advanced automatic modulation recognition (AMC) (O'Shea et al., 2018; Li et al., 2023), jamming pattern recognition (JPR) (Wang, 2019; Wang et al., 2026), and radio frequency (RF)-based unmanned aerial vehicle (UAV) identification (Yu et al., 2025; Cai et al., 2024). EM foundation models, including SpectrumFM and EMind, broaden this scope by learning transferable representations through self-supervised pretraining (Zhou et al., 2025; Luo et al., 2025). MERLIN and PReD further connect signal representations with LLMs for question answering and strategy generation (Shen et al., 2026; Han et al., 2026). However, representative waveform-based models are primarily trained and evaluated on segments containing hundreds to a few thousand in-phase and quadrature (IQ) samples; these evaluations do not establish their ability to understand million-sample raw-IQ recordings (Zhou et al., 2025; Luo et al., 2025; Shen et al., 2026). RF-GPT explores

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Table 1: Comparison of capabilities between existing EM models and LEMMO.

Model	Raw IQ Encoding	Long IQ Sequence	Natural Language QA	Detailed Signal Description	UAV RF Recognition
SpectrumFM	✓	✗	✗	✗	✗
EMind	✓	✗	✗	✗	✗
RF-GPT	✗	✗	✓	✓	✗
PReD	✗	✗	✓	✗	✗
MERLIN	✓	✗	✓	✗	✗
LEMMO	✓	✓	✓	✓	✓

wideband understanding through spectrograms, but magnitude spectrograms do not explicit retain phase information, and their time-frequency detail depends on image resolution (Zou et al., 2026a).

We study instruction-driven understanding directly from continuous wideband IQ recordings spanning multiple frequency channels at a fixed receiver center frequency. The temporal locations and frequency bands of signals of interest are unknown in advance, so the model cannot rely on prior target-specific cropping, bandpass filtering, or downsampling. This setting presents two coupled challenges. First, high sampling rates produce long inputs even over brief intervals: a 10 ms recording at 100 MS/s contains one million complex samples. Processing these inputs requires efficient encoding that preserves local waveform characteristics and integrates context across the recording. Second, the evidence needed from a recording depends on the instruction. Recognition requires category-discriminative features, whereas detailed description requires distinguishing signal components, associating attributes with them, and describing their relationships. The signal-to-language interface must therefore both select relevant evidence and organize it for the requested task.

To address these challenges, we introduce LEMMO, a framework that combines hierarchical waveform processing with instruction-conditioned evidence retrieval. A causal linear-attention EM encoder extracts signal features, and a Temporal Resampler reduces the sequence length for recordings exceeding one millio complex samples. The resampler compresses encoded subsequences by a factor of 64, concatenates them in temporal order, and integrates their context through bidirectional linear attention; shorter recordings bypass resampling. This design supports inputs of up to eight million complex samples. The EM-text interaction module then conditions learnable queries on the instruction before they attend to signal features, making evidence selection instruction-dependent. The retrieved representations are projected into a pretrained LLM's embedding space for response generation.

We further tailor the query organization and supervision to the different requirements of recognition and description. **Q16 Recognition** uses 16 general-purpose queries trained jointly for language generation and classification. **Q128 Description** uses 128 queries organized into global, object-feature, and relation groups, together with physical position information and auxiliary factual supervision. This structure organizes evidence at the scene, object, and relation levels, while the auxiliary objectives supervise the representations used for generation without passing fact-head predictions to the LLM.To support training and evaluation, we construct and release a unified corpus of real-world and synthetic IQ recordings, including LEMMO-ShortQA for recognition and LEMMO-LongDesc for detailed description. Across the recognition tasks evaluated in EM-Bench and LEMMO-ShortQA, LEMMO achieves higher mean accuracy than Merlin-LoRA.

Our contributions are threefold:

- We combine linear-attention waveform encoding, temporal compression, and cross-subsequence contextualization to support inputs of up to eight million complex samples.
- We condition learnable queries on instructions before signal retrieval and develop two complementary configurations: compact recognition queries with discriminative supervision, and structured description queries with auxiliary factual supervision.
- We release real-world and synthetic raw-IQ recordings for representation learning and language supervision, including LEMMO-ShortQA and LEMMO-LongDesc for recognition and detailed descriptions of signal components, attributes, and relationships.

2 Related Work

Research on EM signal understanding follows two directions: EM foundation models learn transferable signal representations, while EM-language models connect signal with natural language.

2.1 EM Foundation Models

EM foundation models use self-supervised learning to extract transferable representations from unlabeled signal data (Shi et al., 2023; Kong et al., 2023; Xiao et al., 2024; Davaslioglu et al., 2022). Their inputs and learning objectives vary. RIS-MAE applies masked reconstruction to raw IQ signals, while IQFM uses contrastive learning on IQ streams (Liu et al., 2025b; Mashaal & Abou-Zeid, 2026). Other efforts, including SpectrumFM, 6G WavesFM, and WiFo, explore representation learning for spectrum management, communications, sensing, localization, and channel prediction (Zhou et al., 2025; Aboulfotouh et al., 2025; Liu et al., 2025a). Recent studies also improve the efficiency and flexibility of signal encoding. Patch-based tokenization reduces sequence length, facilitating the processing of inputs containing thousands of samples (Cheraghinia et al., 2025). EMind introduces length-adaptive multi-signal packing to accommodate heterogeneous EM inputs (Luo et al., 2025). These approaches broaden the range of signals that a common encoder can process.

However, representation learning alone does not provide instruction-driven interaction or natural-language description. Moreover, representative models are evaluated primarily on inputs containing a few thousand samples (Luo et al., 2025; Cheraghinia et al., 2025), which cover only brief observation intervals at high sampling rates. Such intervals may exclude recurring bursts, frequency changes, or relationships between temporally separated events. LEMMO addresses this limitation through hierarchical processing of raw IQ recordings containing up to eight million samples, coupled with an instruction-conditioned language interface.

2.2 EM-Language Models

EM-language models integrate signal information with LLMs in different ways. RadioLLM uses prompt and token reprogramming for radio-signal processing (Chen et al., 2025), while Wireless-LLM fine-tunes LLMs for wireless communication tasks (Shao et al., 2024). MERLIN jointly learns signal representations and adapts a pretrained LLM for EM perception and strategy generation (Shen et al., 2026).

Other approaches transform signals into visual or symbolic representations. Seeing Radio and RF-GPT use image-based signal representations for recognition and text generation (Zou et al., 2026b;a). PReD integrates multiple signal views, including waveforms, spectra, time-frequency representations, and constellation diagrams, to support EM perception, recognition, and decision-making (Han et al., 2026). DiSC-AMC discretizes numerical signal features into symbolic tokens and combines them with nearest-neighbor retrieval for modulation classification (Rostami et al., 2026).

These approaches enable language-based EM processing but differ in the signal information they preserve. Visual representations capture selected signal characteristics, with fidelity determined by the transformation and rendering resolution. In particular, magnitude spectrograms do not explicitly retain phase information. Raw IQ recordings preserve both sampled amplitude and phase information (O'Shea & West, 2016; O'Shea et al., 2018; Davaslioglu et al., 2022), motivating the development of signal-to-language mappings directly from IQ inputs.

Comprehensive EM understanding requires models to process long recordings and generate detailed descriptions of their content. Several waveform-based approaches primarily process short signal segments and focus on classification or concise responses (Chen et al., 2025; Shen et al., 2026; Rostami et al., 2026). Although image-based approaches also support text generation, describing long raw-IQ recordings requires identifying multiple signal components, associating attributes with each component, and characterizing their temporal and spectral relationships. LEMMO addresses these requirements by combining long-recording processing with instruction-conditioned signal retrieval. Its structured queries and auxiliary factual supervision organize the retrieved evidence into scene-level properties, object attributes, and cross-object relationships for description generation.

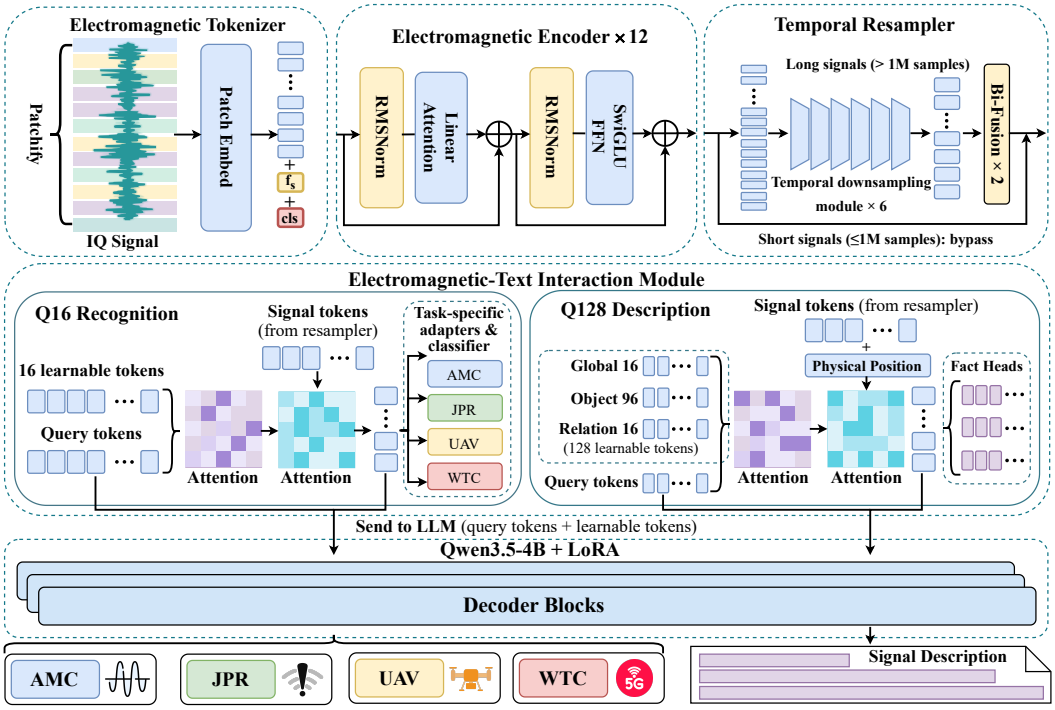


Figure 1: Overview of LEMMO. Raw IQ signals are encoded and temporally compressed for long recordings. Instruction-conditioned queries retrieve signal evidence for a LoRA-adapted LLM, supporting recognition with 16 queries and detailed description with 128 structured queries.

3 METHOD

3.1 OVERVIEW

LEMMO maps raw EM waveforms to natural-language responses through an instruction-conditioned signal interface. Given an IQ recording, its sampling rate, and a natural-language instruction, the model retrieves relevant signal evidence and projects it into the embedding space of a language model. As shown in Figure 1, LEMMO comprises an EM tokenizer, an EM encoder, a temporal resampler, an EM-text interaction module, and a Qwen3.5-4B language backbone. The tokenizer and encoder extract patch-level representations directly from IQ samples. For long recordings, the temporal resampler combines local compression with global information exchange. The interaction module then conditions learnable queries on the textual instruction before retrieving signal evidence, making the signal-to-language mapping dependent on the requested task.

3.2 EM TOKENIZER AND ENCODER

Raw IQ tokenization. The EM tokenizer jointly normalizes the I and Q channels using a common peak-amplitude scale, placing recordings from different wireless technologies on the same scale while preserving relative amplitudes within each recording. The normalized sequence is divided into non-overlapping patches of 16 IQ samples, each mapped to a signal token and augmented with a sinusoidal positional encoding to represent patch order. A sampling-rate token and a learnable CLS token are added to the token sequence to encode the acquisition rate and provide a dedicated representation for global information, respectively.

Efficient waveform encoding. The EM encoder comprises 12 causal linear-attention blocks, with computational cost approximately linear in token-sequence length to support efficient encoding of high-rate IQ signals. We pretrain the encoder using masked waveform reconstruction, in which randomly selected signal patches are masked and reconstructed from the unmasked context. This training encourages the encoder to learn waveform structure before joint EM-language training.

3.3 TEMPORAL RESAMPLER

Even with efficient encoding, long IQ recordings produce too many tokens for instruction-conditioned retrieval. The temporal resampler addresses this bottleneck by compressing the encoded sequence and integrating context across subsequences.

For recordings containing at most one million IQ samples, the temporal resampler is bypassed, and the EM encoder outputs are passed directly to the EM-text interaction module. For longer recordings, the encoded sequence is partitioned into consecutive subsequences, each corresponding to one million IQ samples, with a shorter final subsequence for any remaining samples. Each encoded subsequence is independently processed by a six-stage temporal downsampling module, reducing its token length by a factor of 64. The compressed subsequences are concatenated in their original order and processed by two bidirectional linear-attention layers to integrate context across subsequences. The resulting features are then passed to the EM-text interaction module.

3.4 EM-TEXT INTERACTION MODULE

Different tasks require different evidence from the same recording. The EM-text interaction module selects this evidence through two successive cross-attention operations: learnable queries first attend to the task instruction and then retrieve signal features. The resulting representations provide a compact, instruction-conditioned input to the LLM. This module operates in two configurations according to the tasks: Q16 Recognition uses 16 queries for short-answer recognition tasks, whereas Q128 Description uses 128 structured queries for detailed signal descriptions.

Let T denote the textual instruction-token embeddings obtained from the LLM's embedding layer, and let H denote the signal features output by the Temporal Resampler, or directly by the EM encoder when resampling is bypassed. We write $\mathrm{CrossAttn}(Q, K, V)$ for cross-attention with queries Q, keys K, and values V.

Q16 Recognition. Q16 Recognition maintains a bank of 16 learnable query tokens Q_0, randomly initialized before training. These queries first attend to the instruction embeddings:

$$Q_{\mathrm{inst}} = Q_0 + \mathrm{CrossAttn}(Q_0, T, T). \tag{1}$$

The instruction-conditioned queries then retrieve evidence from the signal features:

$$Q_s = Q_{\mathrm{inst}} + \mathrm{CrossAttn}(Q_{\mathrm{inst}}, H, H). \tag{2}$$

This allows the instruction to guide signal retrieval before passing to the LLM.

The retrieved queries support both answer generation and auxiliary classification. For generation, Q_s is projected into the LLM's embedding space and supplied together with the instruction. For classification, four task-specific residual adapters specialize Q_s for AMC, JPR, UAV recognition, and wireless-technology classification (WTC). Each adapter consists of layer normalization, a linear layer, and a GELU activation within a residual connection:

$$Q^t = Q_s + \mathrm{GELU}\left(\mathrm{Linear}^t\left(\mathrm{LN}^t(Q_s)\right)\right), \tag{3}$$

where $t \in \{\mathrm{AMC}, \mathrm{JPR}, \mathrm{UAV}, \mathrm{WTC}\}$. The adapted representations are passed to the corresponding task-specific classifiers. This auxiliary path encourages the retrieved queries to retain category-discriminative information. We train the whole Q16 module with the following loss:

$$\mathcal{L}_{\mathrm{Q16}} = \mathcal{L}_{\mathrm{LM}} + \lambda_{\mathrm{cls}}\mathcal{L}_{\mathrm{cls}}, \tag{4}$$

where $\mathcal{L}_{\mathrm{LM}}$ is the autoregressive cross-entropy loss on reference answer tokens, $\mathcal{L}_{\mathrm{cls}}$ is the cross-entropy loss between the relevant task-specific classifier's predicted distribution and the ground-truth class label, and λ_{cls} controls the auxiliary loss weight.

Q128 Description. Q128 Description uses 128 learnable queries organized into 16 global queries, 96 object queries, and 16 relation queries. Each signal segment is treated as a signal object. Global queries represent scene-level properties and object counts; object queries form 16 slots with six queries per object; and relation queries capture cross-object context.

The queries follow the same instruction-conditioning and signal-retrieval procedure as Q16 Recognition. To support temporal interpretation, we augment each signal feature h_i with positional information before retrieval:

$$\tilde{h}_i = h_i + \mathrm{PosEnc}(\rho_i), \tag{5}$$

where ρ_i contains the token's position in the full recording, its physical time derived from the sampling rate, and its subsequence index. The resulting sequence $\tilde{H}$ replaces H in the signal-retrieval cross-attention.

The retrieved queries are projected into the LLM's embedding space for description generation. Auxiliary fact heads supervise the same representations by predicting scene-level properties, object counts, object-to-signal associations, signal attributes, and temporal or spectral relations. These heads are used only during training, and their predictions are not supplied to the LLM. The training loss of this module is

$$\mathcal{L}_{\mathrm{Q128}} = \mathcal{L}_{\mathrm{LM}} + \sum_{r \in \mathcal{A}(a)} \lambda_r \mathcal{L}_{\mathrm{fact}}^{(r)}, \tag{6}$$

where $\mathcal{L}_{\mathrm{LM}}$ is the autoregressive cross-entropy loss on reference description tokens and $\mathcal{A}(a)$ denotes the fact types annotated in reference description a. Each auxiliary loss uses cross-entropy for categorical targets or smooth L1 loss for numerical targets, weighted by λ_r. We note that unannotated fact types incur no auxiliary loss during training.

3.5 Supervised Fine-Tuning of LLM

The signal representations Q_s produced by the EM-text interaction module are projected into the embedding space of Qwen3.5-4B and combined with the instruction embeddings T to form the LLM input. The LLM then generates a response conditioned on the instruction and the retrieved signal evidence. We adapt Qwen3.5-4B to EM tasks using rank-16 LoRA by minimizing the autoregressive cross-entropy loss on reference response tokens. The base LLM parameters remain frozen, and only its LoRA parameters are updated.

4 Experiments

4.1 Datasets and Experimental Setup

Datasets. Table 2 summarizes the four datasets used in our experiments. We construct LEMMO-ShortQA and LEMMO-LongDesc to support EM understanding tasks ranging from signal recognition to detailed descriptions of long IQ recordings (Tekbıyık et al., 2020; O'Shea et al., 2016; O'Shea & West, 2016; Schmidt et al., 2019; Luo et al., 2025; Swinney & Woods, 2021; Shi et al., 2025; Subray, 2023; Fontaine et al., 2020; 2019; Liu et al., 2017; Yu et al., 2024a; Ren et al., 2025). LEMMO-ShortQA provides short-answer supervision for four tasks: AMR, JPR, UAV recognition, and WTC. It includs 45 classes, with input lengths ranging from 128 to 5,000,000 IQ samples. LEMMO-LongDesc supports detailed signal description, pairing long IQ recordings with natural-language instructions and reference descriptions. It includes 4G LTE, 5G NR, basic-modulation, ISM, and UAV signals, with input lengths ranging from approximately 400,000 to 8,000,000 IQ samples.

We also use EM-134K and EM-Bench (Shen et al., 2026). EM-134K provides 134,107 public signal–language pairs for instruction tuning. EM-Bench contains 14 tasks with 300 examples per task, comprising ten perception tasks and four strategy-generation tasks. EM-Bench is used exclusively for evaluation.

Baselines and evaluation. We compare LEMMO with MERLIN-LoRA and four general-purpose LLMs on EM-Bench, LEMMO-ShortQA, and LEMMO-LongDesc. Note that the released MERLIN checkpoint is obtained by fully fine-tuning the Qwen3-4B-Instruct-2507 backbone. Although this adaptation improves EM-task performance, it is accompanied by a substantial decline in general knowledge performance: MMLU accuracy decreases from 57.41% for the original backbone to 30.00% for MERLIN. To support EM adaptation while preserving general knowledge, we construct and train MERLIN-LoRA as a baseline. MERLIN-LoRA freezes the Qwen3-4B-Instruct-2507 backbone and applies rank-16 LoRA for language-model adaptation. In addition to evaluating

Table 2: Dataset summary. Each split reports the number of IQ recordings / instruction–response pairs. A dash indicates an unused split. IQ length is measured in complex samples, with M denoting one million.

Dataset	Train	Validation	Test	IQ length
LEMMO-ShortQA	3,832 / 15,328	870 / 3,480	846 / 3,384	128–5M
LEMMO-LongDesc	16,103 / 165,328	856 / 8,813	652 / 802	0.4M–8M
EM-134K	95,679 / 120,829	10,587 / 13,278	–	512/1,024
EM-Bench	–	–	3,370 / 4,200	512/1,024

EM tasks, we assess general knowledge retention on MMLU by comparing MERLIN and MERLIN-LoRA with their Qwen3-4B-Instruct-2507 backbone, and LEMMO with its Qwen3.5-4B backbone.

4.2 EM-Bench Performance and General Knowledge Retention

Table 3: Performance on the 14 EM-Bench tasks. Perception tasks report accuracy (Acc., %), and strategy-generation tasks report ROUGE-L F1 (R-L, scaled to 0–100). Averages are computed across tasks within each group. The highest score in each row is bold.

Task	Metric	**LEMMO**	MERLIN -LoRA	GPT-6 Luna	DeepSeek V4.1-Flash	Gemini-3.5 Flash-Lite	Qwen3.8 Flash
Perception							
CJR	Acc.	**83.67**	72.33	13.00	11.67	15.33	21.33
MOD	Acc.	**36.00**	27.00	0.00	1.33	9.33	10.33
Radar bandwidth	Acc.	**58.33**	27.00	0.00	0.00	0.33	1.33
Radar duty cycle	Acc.	**54.33**	30.67	1.67	0.67	14.67	11.00
Radar NP	Acc.	**82.00**	81.00	9.33	1.00	23.00	18.33
Radar PRF	Acc.	**41.33**	26.33	0.00	0.00	4.00	3.00
Radar pulse width	Acc.	**37.33**	30.33	0.00	0.00	7.00	3.33
PI	Acc.	**79.67**	71.00	3.33	10.00	7.67	16.00
RJR	Acc.	**73.67**	45.67	24.67	12.67	19.33	27.67
SD	Acc.	46.67	24.00	27.00	36.33	**56.67**	56.00
Perception average	Acc.	**59.30**	43.53	7.90	7.37	15.73	16.83
Generation							
Anti-CJ	R-L	50.68	**51.14**	0.22	1.31	7.81	8.27
Anti-RJ	R-L	**43.85**	35.12	0.64	1.30	7.64	9.16
CJS	R-L	39.84	**44.64**	1.32	3.77	8.36	7.04
RJS	R-L	**56.17**	42.14	1.35	5.60	10.28	9.37
Generation average	R-L	**47.64**	43.26	0.88	3.00	8.52	8.46

Table 3 shows that LEMMO achieves the highest scores on 11 of the 14 EM-Bench tasks. It outperforms MERLIN-LoRA on all ten perception tasks, achieving 59.30% mean accuracy (+15.77 percentage points) and ranking first on nine tasks. For strategy generation, LEMMO leads on Anti-RJ and RJS and achieves the highest mean ROUGE-L F1 of 47.64, exceeding MERLIN-LoRA by 4.38 points. General-purpose LLMs perform substantially worse on most tasks, although Gemini-3.5-Flash-Lite leads on SD. These results demonstrate LEMMO's effectiveness in both EM perception and strategy generation.

Table 4 reports MMLU accuracy before and after EM adaptation. LEMMO achieves 64.98% overall accuracy, compared with 64.34% for its Qwen3.5-4B backbone. MERLIN-LoRA similarly preserves its backbone's accuracy (57.55% versus 57.41%), whereas the released MERLIN checkpoint scores 30.00%, a decline of 27.41 percentage points. These findings show that LEMMO achieves strong EM-task performance while preserving general knowledge.

Table 4: MMLU accuracy (%) for EM-adapted models and their original language backbones. MERLIN and MERLIN-LoRA use Qwen3-4B-Instruct-2507; LEMMO uses Qwen3.5-4B. The highest score in each column is bold.

Model	Overall	STEM	Humanities	Social Sciences	Other
MERLIN	30.00	28.45	27.97	32.95	31.70
MERLIN-LoRA	57.55	53.60	48.33	70.26	62.92
Qwen3-4B-Instruct-2507	57.41	50.02	53.41	68.38	60.12
Qwen3.5-4B	64.34	58.77	**59.70**	71.86	69.58
LEMMO	**64.98**	**60.04**	59.49	**72.64**	**70.71**

4.3 EM Recognition on LEMMO-ShortQA

Table 5 reports open-ended recognition accuracy on LEMMO-ShortQA. LEMMO achieves the highest accuracy on all four tasks, with a mean of 70.05%, exceeding MERLIN-LoRA by 14.31%. It improves AMC, JPR, and WTC by 9.71, 7.77, and 7.89 %, respectively. The largest gain occurs in UAV recognition, where accuracy increases from 41.88% to 73.75% (+31.87 %). All four general-purpose LLMs obtain substantially lower accuracy across the tasks.

Table 5: Recognition accuracy (%) on the LEMMO-ShortQA test set. Average denotes the unweighted mean across the four tasks. The highest score in each column is bold.

Model	AMC	JPR	UAV	WTC	Average
GPT-6 Luna	0.32	4.00	0.00	0.08	1.10
DeepSeek-V4.1-Flash	0.81	3.12	0.00	1.95	1.47
Gemini-3.5-Flash-Lite	2.35	10.29	0.00	2.42	3.77
Qwen3.8-Flash	4.14	3.48	0.00	4.14	2.94
MERLIN-LoRA	60.01	61.47	41.88	59.61	55.74
LEMMO	**69.72**	**69.24**	**73.75**	**67.50**	**70.05**

Table 6 summarizes the input lengths for each task. UAV recognition uses the longest recordings, each containing five million IQ samples, and also exhibits the largest improvement over MERLIN-LoRA. Together with the gains on shorter inputs, this result demonstrates LEMMO's effectiveness across recognition tasks with widely varying signal lengths.

Table 6: Input IQ lengths for each LEMMO-ShortQA task, measured in complex samples.

AMC	JPR	UAV	WTC
128–1,024 samples	128–16,368 samples	5,000,000 samples	4,096 samples

4.4 Long-Recording Description on LEMMO-LongDesc

We evaluate two tasks on the LEMMO-LongDesc test set: multi-signal scene overview and per-signal detailed description. Table 7 reports both text-overlap and fact-level metrics. ROUGE-1 F1 (R-1), ROUGE-2 F1 (R-2), ROUGE-L F1 (R-L), BLEU-4, and Token F1 measure agreement with reference descriptions. Category F1 evaluates the recovery of categorical facts, such as modulation type, while Value F1 evaluates numerical facts, including signal counts, start and end times, and center frequencies.

LEMMO leads on all seven metrics for multi-signal scene overview. Compared with MERLIN-LoRA, it improves ROUGE-L F1 by 4.66 points and Token F1 by 4.49 points. The gains are larger for factual content: Category F1 increases from 39.76 to 44.83 (+5.07 points), and Value F1 increases from 36.10 to 44.83 (+8.73 points). For per-signal detailed description, LEMMO achieves the highest scores on six of seven metrics. Its main improvements are in Category F1, which rises from 34.43 to 41.42 (+6.99 points), and Value F1, which rises from 19.61 to 24.06 (+4.45 points).

Table 7: Description performance on the LEMMO-LongDesc test set. All scores are scaled to 0–100. The highest score for each metric within each task is bold.

	Text Metrics					Fact Metrics	
Model	R-1	R-2	R-L	BLEU-4	Token F1	Category F1	Value F1
Multi-signal scene overview task							
GPT-6 Luna	22.79	6.26	11.49	3.86	23.51	0.00	0.00
DeepSeek-V4.1-Flash	21.88	5.95	10.36	2.64	23.01	0.00	0.50
Gemini-3.5-Flash-Lite	20.32	5.33	11.25	0.93	21.89	0.10	0.00
Qwen3.8-Flash	20.62	5.50	10.69	1.99	22.02	1.20	0.30
MERLIN-LoRA	72.76	61.37	66.17	49.43	70.31	39.76	36.10
LEMMO	**76.45**	**65.43**	**70.83**	**50.94**	**74.80**	**44.83**	**44.83**
Per-signal detailed description task							
GPT-6 Luna	31.78	8.62	13.83	4.79	32.02	0.00	0.48
DeepSeek-V4.1-Flash	28.22	8.21	12.53	3.60	28.94	0.27	2.86
Gemini-3.5-Flash-Lite	26.48	7.52	14.47	2.97	27.51	0.00	0.00
Qwen3.8-Flash	26.05	7.53	11.81	2.98	27.16	0.75	0.03
MERLIN-LoRA	68.41	54.16	64.71	**43.63**	66.83	34.43	19.61
LEMMO	**68.51**	**54.82**	**65.11**	42.96	**67.77**	**41.42**	**24.06**

Table 8: Effect of input IQ length on fact metrics. The same questions are evaluated at each length.

Input IQ length	4K	64K	256K	1M	4M	4K→4M
Category F1	42.13	49.31	61.69	63.08	**68.45**	**+26.32**
Value F1	13.03	24.76	24.77	20.00	**35.24**	**+22.21**

Text-overlap gains are smaller, while BLEU-4 decreases by 0.67 points. General-purpose LLMs obtain substantially lower scores on both tasks, particularly on the fact-level metrics. These results demonstrate LEMMO's effectiveness in describing long recordings, with consistent improvements in the recovery of categorical and numerical signal facts.

4.5 Effect of Input Signal Length

Table 8 evaluates factual prediction using the same instructions with input lengths ranging from 4K to 4M IQ samples. Category F1 increases consistently with input length, rising from 42.13 to 68.45. Value F1 also peaks at 4M samples, increasing from 13.03 to 35.24, although it decreases at 1M samples. Relative to 4K inputs, 4M inputs improve Category F1 and Value F1 by 26.32 and 22.21 points, respectively. These results indicate that LEMMO can use longer signal observations to improve factual prediction, highlighting the effectiveness of long-recording EM modeling.

5 Conclusion

We introduced LEMMO, a framework for instruction-driven EM understanding from raw IQ recordings of up to eight million samples. LEMMO combines hierarchical long-sequence processing with an EM–text interface that conditions learnable queries on instructions before retrieving signal evidence. Its complementary query designs support both recognition and detailed description: compact queries receive discriminative supervision, while structured queries organize scene, object, and relation information with auxiliary factual supervision. We also constructed LEMMO-ShortQA and LEMMO-LongDesc to support training and evaluation across these tasks. LEMMO significantly outperforms MERLIN-LoRA on average across EM-Bench, on all four LEMMO-ShortQA tasks, and on most LEMMO-LongDesc metrics, while retaining its LLM backbone's MMLU performance. Input-length experiments further demonstrate improved factual prediction with longer observations, highlighting the value of combining long raw-IQ processing with instruction-conditioned, structured evidence retrieval for EM understanding.

AI USE STATEMENT

We used generative AI tools to assist with writing code, monitoring experiment progress, and generating question–answer text for LEMMO-ShortQA and LEMMO-LongDesc. We also used these tools to polish the language of the manuscript. All AI-generated code, dataset text, manuscript edits, and other artifacts were manually reviewed. We take full responsibility for the final content of this work, including its text, code, datasets, and scientific claims.

REPRODUCIBILITY STATEMENT

To support reproducibility, Appendix A details the model architecture and IQ preprocessing. Appendix B reports the training stages, objectives, hyperparameters, and checkpoint selection criteria. Appendix C describes the dataset sources and composition, data splits, task definitions, and evaluation protocols for LEMMO-ShortQA, LEMMO-LongDesc, EM-134K, and EM-Bench. Our code and datasets are available anonymously at `https://anonymous.4open.science/r/20B5` and `https://anonymous-hf.com/a/nligi354i8tn/`, respectively.

REFERENCES

Ahmed Aboulfotouh, Elsayed Mohammed, and Hatem Abou-Zeid. 6G WavesFM: A foundation model for sensing, communication, and localization. *IEEE Open Journal of the Communications Society*, 2025.

Zhenxin Cai, Yu Wang, Qi Jiang, Guan Gui, and Jin Sha. Toward intelligent lightweight and efficient UAV identification with RF fingerprinting. *IEEE Internet of Things Journal*, 11(15):26329–26339, 2024. doi: 10.1109/JIOT.2024.3395466.

Shuai Chen, Yong Zu, Zhixi Feng, Shuyuan Yang, and Mengchang Li. RadioLLM: Introducing large language model into cognitive radio via hybrid prompt and token reprogrammings. *arXiv preprint arXiv:2501.17888*, 2025.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024.

Mohammad Cheraghinia, Eli De Poorter, Jaron Fontaine, Merouane Debbah, and Adnan Shahid. A unified foundation model for wireless technology recognition and localization. *arXiv preprint arXiv:2505.19390*, 2025.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

Kemal Davaslioglu, Serdar Boztaş, Mehmet Can Ertem, Yalin E Sagduyu, and Ender Ayanoglu. Self-supervised RF signal representation learning for NextG signal classification with deep learning. *IEEE Wireless Communications Letters*, 12(1):65–69, 2022.

Yihong Dong, Xiaohan Jiang, Huaji Zhou, Yun Lin, and Qingjiang Shi. SR2CNN: Zero-shot learning for signal recognition. *IEEE Transactions on Signal Processing*, 69:2316–2329, 2021. doi: 10.1109/TSP.2021.3070186.

Jaron Fontaine, Erika Fonseca, Adnan Shahid, Maicon Kist, Luiz A. DaSilva, Ingrid Moerman, and Eli De Poorter. Towards low-complexity wireless technology classification across multiple environments. *Ad Hoc Networks*, 91:101881, 2019. doi: 10.1016/j.adhoc.2019.101881. URL `https://doi.org/10.1016/j.adhoc.2019.101881`.

Jaron Fontaine, Adnan Shahid, Robbe Elsas, Amina Seferagic, Ingrid Moerman, and Eli De Poorter. Multi-band sub-GHz technology recognition on NVIDIA's Jetson Nano. In *2020 IEEE 92nd Vehicular Technology Conference (VTC2020-Fall)*, pp. 1–7. IEEE, 2020.

doi: 10.1109/VTC2020-Fall49728.2020.9348566. URL https://doi.org/10.1109/VTC2020-Fall49728.2020.9348566.

Zehua Han, Jing Xiao, Yiqi Duan, Mengyu Xiang, Yuheng Ji, Xiaolong Zheng, Chenghanyu Zhang, Zhendong She, Junyu Shen, Dingwei Tan, et al. PReD: An LLM-based foundation multimodal model for electromagnetic perception, recognition, and decision. *arXiv preprint arXiv:2603.28183*, 2026.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, et al. Language is not all you need: Aligning perception with language models. *Advances in Neural Information Processing Systems*, 36: 72096–72109, 2023.

Weisi Kong, Xun Jiao, Yuhua Xu, Bolin Zhang, and Qinghai Yang. A transformer-based contrastive semi-supervised learning framework for automatic modulation recognition. *IEEE Transactions on Cognitive Communications and Networking*, 9(4):950–962, 2023.

Taotao Li, Zhenyu Wen, Yang Long, Zhen Hong, Shilian Zheng, Li Yu, Bo Chen, Xiaoniu Yang, and Ling Shao. The importance of expert knowledge for automatic modulation open set recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):13730–13748, 2023. doi: 10.1109/TPAMI.2023.3294505.

Boxun Liu, Shijian Gao, Xuanyu Liu, Xiang Cheng, and Liuqing Yang. WiFo: Wireless foundation model for channel prediction. *Science China Information Sciences*, 68(6):162302, 2025a.

Wei Liu, Merima Kulin, Tarik Kazaz, Adnan Shahid, Ingrid Moerman, and Eli De Poorter. Wireless technology recognition based on RSSI distribution at sub-Nyquist sampling rate for constrained devices. *Sensors*, 17(9):2081, 2017. doi: 10.3390/s17092081. URL https://doi.org/10.3390/s17092081.

Yunfei Liu, Mingxuan Liu, Wupeng Xie, Xinzhu Liu, Wenxue Liu, Yangang Sun, Xin Qiu, Cui Yuan, and Jinhai Li. RIS-MAE: A self-supervised modulation classification method based on raw IQ signals and masked autoencoder. *arXiv preprint arXiv:2508.00274*, 2025b.

Luqing Luo, Wenjin Gui, Yunfei Liu, Ziyue Zhang, Yunxi Zhang, Fengxiang Wang, Zonghao Guo, Zizhi Ma, Xinzhu Liu, Hanxiang He, et al. EMind: A foundation model for multi-task electromagnetic signals understanding. *arXiv preprint arXiv:2508.18785*, 2025.

Omar Mashaal and Hatem Abou-Zeid. IQFM–a wireless foundation model for I/Q streams in ai-native 6G. *IEEE Open Journal of the Communications Society*, 2026.

Timothy J O'Shea and Nathan West. Radio machine learning dataset generation with GNU radio. In *Proceedings of the GNU radio conference*, volume 1, 2016.

Timothy J. O'Shea, Johnathan Corgan, and T. Charles Clancy. Convolutional radio modulation recognition networks. In *Engineering Applications of Neural Networks*, volume 629 of *Communications in Computer and Information Science*, pp. 213–226. Springer, 2016. doi: 10.1007/978-3-319-44188-7_16. URL https://doi.org/10.1007/978-3-319-44188-7_16.

Timothy James O'Shea, Tamoghna Roy, and T Charles Clancy. Over-the-Air deep learning based radio signal classification. *IEEE Journal of Selected Topics in Signal Processing*, 12(1):168–179, 2018.

Junyu Ren, Ningning Yu, Chengwei Zhou, Zhiguo Shi, and Jiming Chen. DroneRFb-DIR: An RF signal dataset for non-cooperative drone individual identification. *Journal of Electronics & Information Technology*, 47(3):573–581, 2025. doi: 10.11999/JEIT240804. URL https://doi.org/10.11999/JEIT240804.

Mohammad Rostami, Atik Faysal, Reihaneh Gh. Roshan, Huaxia Wang, Nikhil Muralidhar, and Yu-Dong Yao. Large language models can perform automatic modulation classification via discretized self-supervised candidate retrieval, 2026. URL https://arxiv.org/abs/2510.00316.

Malte Schmidt, Dimitri Block, and Uwe Meier. CRAWDAD dataset owl/interference (v. 2019-02-12). CRAWDAD, hosted by IEEE DataPort, 2019. URL https://ieee-dataport.org/open-access/crawdad-owl-init.

Jiawei Shao, Jingwen Tong, Qiong Wu, Wei Guo, Zijian Li, Zehong Lin, and Jun Zhang. WirelessLLM: Empowering large language models towards wireless intelligence. *Journal of Communications and Information Networks*, 9(2):99–112, 2024. doi: 10.23919/JCIN.2024.10582827.

Junyu Shen, Zhendong She, Chenghanyu Zhang, Yuchuang Sun, Luqing Luo, Dingwei Tan, Zonghao Guo, Bo Guo, Zehua Han, Wupeng Xie, Yaxin Mu, Peng Zhang, Peipei Li, Fengxiang Wang, Yangang Sun, and Maosong Sun. MERLIN: Building low-snr robust multimodal LLMs for electromagnetic signals. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8653–8663, June 2026.

Rui Shi, Xiaodong Yu, Shengming Wang, Yijia Zhang, Lu Xu, Peng Pan, and Chunlai Ma. RFUAV: A benchmark dataset for unmanned aerial vehicle detection and identification, 2025. URL https://arxiv.org/abs/2503.09033.

Yunhao Shi, Hua Xu, Yue Zhang, Zisen Qi, and Dan Wang. GAF-MAE: A self-supervised automatic modulation classification method based on gramian angular field and masked autoencoder. *IEEE Transactions on Cognitive Communications and Networking*, 10(1):94–106, 2023.

Siddhartha Subray. Real-world wireless communication dataset: IEEE 802.11ax, LTE, and 5G-NR signals. IEEE DataPort, 2023. URL https://doi.org/10.21227/b82j-sy57.

Carolyn J. Swinney and John C. Woods. Raw IQ dataset for GNSS GPS jamming signal classification. Zenodo, 2021. URL https://doi.org/10.5281/zenodo.4629685. Version 1.

Kürşat Tekbıyık, Ali Rıza Ekti, Ali Görçin, Güneş Karabulut Kurt, and Cihat Keçeci. Robust and fast automatic modulation classification with CNN under multipath fading channels. In *2020 IEEE 91st Vehicular Technology Conference (VTC2020-Spring)*, 2020. doi: 10.1109/VTC2020-Spring48590.2020.9128408. URL https://doi.org/10.1109/VTC2020-Spring48590.2020.9128408.

Pengyu Wang, Zhaocheng Wang, Tianqi Mao, Weijie Yuan, Haijun Zhang, and George K. Karagiannidis. Jamming identification with differential transformer for low-altitude wireless networks. *IEEE Journal on Selected Areas in Communications*, 44:3664–3677, 2026. doi: 10.1109/JSAC.2026.3669137.

Yafeng Wang. Recognition of radar active-jamming through convolutional neural networks. *The Journal of Engineering*, 2019(21):7695–7697, 2019. doi: 10.1049/joe.2019.0659.

Chenghong Xiao, Shuyuan Yang, Zhixi Feng, and Licheng Jiao. MCLHN: Toward automatic modulation classification via masked contrastive learning with hard negatives. *IEEE Transactions on Wireless Communications*, 23(10):14304–14319, 2024.

Ningning Yu, Shengjian Mao, Chengwei Zhou, Guowei Sun, Zhiguo Shi, and Jiming Chen. DroneRFa: A large-scale dataset of drone radio frequency signals for detecting low-altitude drones. *Journal of Electronics & Information Technology*, 46(4):1147–1156, 2024a. doi: 10.11999/JEIT230570. URL https://doi.org/10.11999/JEIT230570.

Ningning Yu, Jiajun Wu, Chengwei Zhou, Zhiguo Shi, and Jiming Chen. Open set learning for RF-based drone recognition via signal semantics. *IEEE Transactions on Information Forensics and Security*, 19:9894–9909, 2024b. doi: 10.1109/TIFS.2024.3463535.

Ningning Yu, Jiajun Wu, Chengwei Zhou, Zhiguo Shi, and Jiming Chen. SMNet: Multi-drone detection and classification via monitoring flight control signals on spectrograms. *IEEE Transactions on Cognitive Communications and Networking*, 11(5):3245–3259, 2025. doi: 10.1109/TCCN.2025.3537100.

Weishan Zhang, Yue Wang, Xiang Chen, Zhipeng Cai, and Zhi Tian. Spectrum Transformer: An attention-based wideband spectrum detector. *IEEE Transactions on Wireless Communications*, 23(9):12343–12353, 2024. doi: 10.1109/TWC.2024.3391515.

Fuhui Zhou, Chunyu Liu, Hao Zhang, Wei Wu, Qihui Wu, Tony QS Quek, and Chan-Byoung Chae. SpectrumFM: A foundation model for intelligent spectrum management. *IEEE Journal on Selected Areas in Communications*, 2025.

Hang Zou, Yu Tian, Bohao Wang, Lina Bariah, Samson Lasaulce, Chongwen Huang, and Mérouane Debbah. RF-GPT: Teaching ai to see the wireless world. *arXiv preprint arXiv:2602.14833*, 2026a.

Hang Zou, Bohao Wang, Yu Tian, Lina Bariah, Chongwen Huang, Samson Lasaulce, and Mérouane Debbah. Seeing Radio: From zero RF priors to explainable modulation recognition with vision language models. In *2026 IEEE International Conference on Communications Workshops (ICC Workshops)*, pp. 1–6. IEEE, 2026b.

A Detailed Model Structure

This section provides further details on the model components introduced in Section 3. The model divides IQ samples into patches and encodes them. It compresses the resulting features when the recording exceeds one million samples, then passes signal information to Qwen3.5-4B. Q16 Recognition and Q128 Description share the same tokenizer and encoder architecture but use different queries and training objectives.

A.1 EM Tokenizer and EM Encoder

IQ normalization and patching. Let $X \in \mathbb{R}^{L\times 2}$ be an IQ recording with L complex samples. Its two columns contain the in-phase (I) and quadrature (Q) components. Both channels are normalized by the same record-level scale, $s(X) = \max_{t,c} |X_{t,c}|$, the maximum absolute value across all I and Q values in the recording.

After normalization, every 16 consecutive complex samples form one non-overlapping patch. Each patch contains 32 real values, which a shared linear layer maps to a 768-dimensional vector. A recording of length L yields $\lfloor L/16 \rfloor$ complete patches; a final segment shorter than 16 samples is omitted. Each independently encoded sequence also receives a learnable CLS token and a sampling-rate token. The sampling-rate token is produced by a learnable multilayer perceptron (MLP) from fixed-frequency features of the sampling rate. Fixed sinusoidal position encoding restarts at the beginning of each sequence.

Multiple short recordings are placed in one computation container with a total capacity of one million IQ samples. The encoder receives the boundaries of each recording and computes the attention separately within each recording, preventing information exchange between recordings. A recording longer than one million samples is divided into consecutive regions. All regions use the same patch embedding layer and EM Encoder.

The EM Encoder has 12 layers. Each layer has a hidden size of 768, 12 attention heads, and a feed-forward hidden size of 3072. The layers use RMSNorm, FLA causal LinearAttention, and a SwiGLU feed-forward network, with residual connections after the attention and feed-forward sublayers. The patch embedding and encoder parameters are shared across input lengths.

Recordings of at most one million complex IQ samples bypass the Temporal Resampler. Longer recordings are divided into regions of at most one million samples. For each complete region, 62,500 encoded patch tokens are reduced to 977 tokens by six temporal downsampling stages and then pooled to 800 tokens, as detailed in Appendix A.2.

Masked waveform reconstruction. During pretraining, a random subset of patches is masked, and the encoder processes only visible patches. A shared reconstruction decoder with eight layers and a hidden size of 384 inserts learnable mask tokens at the missing positions and predicts the 32 real values of each masked patch.

Before computing the loss, the 32 target values of each masked patch p are standardized using the mean μ_p and variance σ_p^2. The reconstruction loss averages squared error over the 32 values and all

masked patches:

$$\mathcal{L}_{\text{rec}} = \frac{1}{|\mathcal{M}|} \sum_{p \in \mathcal{M}} \frac{1}{32} \left\| \hat{x}_p - \frac{x_p - \mu_p}{\sqrt{\sigma_p^2 + \epsilon}} \right\|_2^2, \qquad \epsilon = 10^{-6}. \tag{7}$$

Here $\mathcal{M}$ is the set of masked patches, x_p is the original patch, and $\hat{x}_p$ is the decoder prediction. The decoder is used for pretraining and is discarded for question answering.

A.2 Temporal Resampler

The Temporal resampler processes only recordings longer than one million IQ samples. The recording is divided in temporal order into regions of at most one million samples. The EM Encoder encodes each region independently, and the resampler then compresses its encoded features.

The Temporal Resampler first applies six temporal downsampling stages, with an overall downsampling factor of 64. Each stage contains a low-pass downsampling path and a gated learnable convolutional path. The low-pass path reduces aliasing, while the convolutional path captures local variation. For a complete region containing one million complex IQ samples, the encoder produces 62,500 patch tokens. The six-stage downsampling stack reduces this sequence to 977 tokens, after which an average-pooling layer reduces it to 800 tokens. A shorter final region receives an output budget proportional to its number of valid patch tokens.

The compressed features are concatenated in the original temporal order. The two bidirectional linear-attention layers then exchange information between regions. Thus, the encoder represents the signal within each region before compression, and information is exchanged between regions after compression.

The compressed features retain their temporal order. At the maximum supported length of eight million samples, eight complete regions yield 6,400 features for the EM-Text Interaction Mechanism.

A.3 EM-Text Interaction Mechanism

This module uses learnable query vectors to extract information from the question and signal for answer generation. The queries first attend to the question tokens and then to the signal features, following Eqs. (1)–(2) for Q16 Recognition and Eqs. (4)–(6) for Q128 Description. Q16 Recognition uses 16 queries, whereas Q128 Description uses 128. Each query output is projected from 768 dimensions to the 2560-dimensional embedding space of Qwen3.5-4B and enters the language model as a separate signal token.

Q16 Recognition. This configuration uses 16 learnable query vectors for AMC, JPR, UAV recognition, and WTC. The queries first attend to the question tokens and then to the signal features, producing the retrieved representations Q_s. These representations are projected into the Qwen3.5-4B embedding space and provided to the LLM together with the question tokens for answer generation.

During supervised training, Q_s also enters an auxiliary classification branch. Each task has a residual adapter consisting of LayerNorm, a $768 \to 64$ linear layer, GELU, and a $64 \to 768$ linear layer, followed by a residual addition. Learned attention weights aggregate the adapted representations into a task representation for the corresponding classifier. The adapters and classifiers provide auxiliary classification supervision during training. At inference, answer generation uses the direct Q_s-to-LLM path without invoking the classification branch.

Q128 Description. This instance uses 128 queries: 16 global queries for information about the full recording, 96 object-feature queries for individual signal objects, and 16 relation queries for interactions between objects. The 96 object-feature queries form 16 groups of six, with one group per object. Query-group, object-index, and feature-type embeddings specify the purpose of each query before it attends to the question.

Q128 also adds three position coordinates to each signal feature: its relative position in the full recording, normalized log physical time computed from the sampling rate, and its position within the long-record region structure. Each coordinate retains its original value and receives sine and

cosine features at eight Fourier frequencies, giving $3 \times (1 + 8 \times 2) = 51$ dimensions in total. A $51 \to 256 \to 768$ projection maps this position vector to the signal-feature dimension. The projected vector is added to the signal feature and followed by LayerNorm. Position information is added before the queries attend to the signal. The resulting 128 query representations support both language generation and auxiliary fact prediction.

Four types of heads provide auxiliary predictions. Scene-level heads predict global categories and counts. Object-level heads predict existence, categories, numeric attributes, and multi-label attributes. A binding head scores associations between object representations and time-ordered signal features. Pairwise heads predict relations between object representations; these heads are separate from the 16 relation queries. Their predictions are used only for training and are not converted into fact tokens for Qwen3.5-4B.

For answer a, let $\mathcal{A}(a)$ be the set of annotated fact types. The training loss combines language-model loss with auxiliary losses for those fact types:

$$\mathcal{L}_{\text{desc}} = \mathcal{L}_{\text{LM}} + \sum_{r \in \mathcal{A}(a)} \lambda_r \mathcal{L}_{\text{fact}}^{(r)}. \tag{8}$$

Here λ_r is the weight of fact type r. Unannotated fact types do not contribute to the auxiliary loss and are not treated as negative examples.

B Training and Implementation Details

This section describes three experimental configurations: Q16 Recognition for LEMMO-ShortQA, Q128 Description for LEMMO-LongDesc, and the direct encoder-to-LLM interface for EM-Bench. The configurations use the same EM Encoder architecture with configuration-specific weights and training stages. All language-generation objectives use causal cross-entropy computed on answer tokens.

B.1 Signal Representation Pretraining

The EM Encoder is pretrained by masked waveform reconstruction on AMC, JPR, UAV, and WTC IQ recordings, sampled in proportions of 20%, 30%, 30%, and 20%, respectively. Both I and Q channels are normalized by the maximum absolute value across the two channels of each recording. The EM Tokenizer maps each patch to a 768-dimensional token. The patch embedding and reconstruction decoder are shared across all recording lengths.

With a 50% mask ratio, the EM Encoder processes visible patch tokens, and an eight-block decoder with hidden width 384 reconstructs the masked patches. Each target patch is standardized using the mean and variance of its 32 real-valued elements. The loss is the mean squared reconstruction error over masked patches.

Pretraining runs for three passes on four H100 GPUs. AdamW uses a peak learning rate of 5×10^{-5}, weight decay 0.01, a 5% warmup, and gradient clipping at norm 1.0. For Q128 Description, the EM Encoder undergoes two further passes of masked reconstruction on 4G LTE, 5G NR, basic-modulation, ISM, and UAV recordings. This stage uses the same EM Tokenizer with patch size 16 and a 50% mask ratio. The peak learning rates are 10^{-5} and 10^{-6} in sequence.

B.2 Training Q16 Recognition

Q16 Recognition is trained on LEMMO-ShortQA, which contains 15,328 training question–answer pairs from 3,832 signals and 3,480 validation pairs from 870 signals. Its four tasks are AMC, JPR, UAV recognition, and WTC. Each signal has four question templates. Training samples are balanced across tasks, source datasets, and labels.

The Temporal Resampler is first trained to compress EM Encoder features. For recordings longer than one million complex IQ samples, the Temporal Resampler reduces the 62,500 patch tokens of each complete one-million-sample region to 977 tokens through six downsampling stages, followed by average pooling to obtain 800 tokens. Two bidirectional fusion layers integrate context across the compressed regions. Training runs for three passes with a peak learning rate of 3×10^{-4}. The

objective combines compressed-feature, cosine, temporal-difference, and masked-fusion losses with weights 1, 0.25, 0.25, and 1, respectively.

Q16 Recognition uses 16 question-conditioned queries in the EM-Text Interaction Mechanism. Q16 Recognition is first trained with the EM Encoder, Temporal Resampler, and Qwen3.5-4B frozen. The final two EM Encoder blocks are then unfrozen for further alignment with Qwen3.5-4B. Language-model adaptation uses rank-16 LoRA with scaling parameter 32 and dropout 0.05. The Qwen3.5-4B base parameters remain frozen. During supervised training, the retrieved query representations Q_s support both answer generation and auxiliary classification. The language-generation branch projects Q_s directly into the LLM embedding space. The classification branch applies task-specific residual adapters with bottleneck width 64, followed by attention pooling and classification. The final joint stage optimizes the answer-generation loss and the auxiliary classification loss:

$$\mathcal{L}_{\text{recog}} = \mathcal{L}_{\text{LM}} + 0.5\,\mathcal{L}_{\text{cls}}. \tag{9}$$

At inference, only the language-generation branch is used.

The final joint stage runs for three passes on four H100 GPUs. The peak learning rates are 10^{-7} for the EM Encoder, 2×10^{-6} for Q16 Recognition, 5×10^{-5} for the task-specific adapters and classifiers, and 5×10^{-6} for Qwen3.5-4B LoRA. AdamW uses $(\beta_1, \beta_2) = (0.9, 0.95)$, weight decay 0.01, a 5% warmup, and gradient clipping at norm 1.0. Checkpoint selection uses label-balanced validation accuracy across the four tasks.

B.3 Training Q128 Description

Q128 Description is trained on LEMMO-LongDesc, which contains 165,328 training question–answer pairs from 16,103 signals and 8,813 validation pairs from 856 signals. Its five signal families are 4G LTE, 5G NR, basic modulation, ISM, and UAV. Question–answer pairs are grouped by physical signal. Each training pass samples five questions per physical signal.

The Temporal Resampler processes recordings containing up to eight million complex IQ samples using the downsampling and average-pooling procedure described in Appendix A.2. Each complete one-million-sample region produces 800 tokens, while a shorter final region receives an output budget proportional to its number of valid patch tokens. The Temporal Resampler is trained for three passes with a peak learning rate of 3×10^{-4} and a 25% feature-mask ratio. The objective combines compressed-feature, cosine, temporal-difference, and masked-fusion losses with weights 1, 0.25, 0.25, and 1, respectively.

Q128 Description uses 16 global queries, 96 object-feature queries, and 16 relation queries in the EM-Text Interaction Mechanism. The object-feature queries form 16 groups of six. Initial alignment proceeds in three stages. Q128 Description is trained first, followed by adaptation of the final two EM Encoder blocks and then training of rank-16 Qwen3.5-4B LoRA. Structured training uses auxiliary fact heads to apply scene-level, object-level, binding, and pairwise-relation supervision to the query representations. Auxiliary fact predictions are not converted to input tokens for Qwen3.5-4B. For answer a, the training objective is

$$\mathcal{L}_{\text{desc}} = \mathcal{L}_{\text{LM}} + \sum_{r \in \mathcal{A}(a)} \lambda_r \mathcal{L}_{\text{fact}}^{(r)}, \tag{10}$$

where $\mathcal{A}(a)$ is the set of fact types annotated for answer a, and λ_r weights the auxiliary loss for fact type r.

During structured training, the final four, final eight, and all twelve EM Encoder blocks are progressively unfrozen. The cross-region fusion layers in the Temporal Resampler are trainable during these stages. The final joint stage also trains Q128 Description, the auxiliary fact heads, and rank-16 Qwen3.5-4B LoRA. The Qwen3.5-4B base parameters remain frozen throughout training.

The final stage uses a peak learning rate of 2×10^{-8} at the top of the EM Encoder, with layer-wise learning-rate decay of 0.85. The other peak learning rates are 5×10^{-8} for the cross-region fusion layers in the Temporal Resampler, 3×10^{-6} for Q128 Description, and 10^{-5} for Qwen3.5-4B LoRA. AdamW uses $(\beta_1, \beta_2) = (0.9, 0.95)$, weight decay 0.05, a 5% warmup, and gradient clipping at norm 1.0. Validation covers all 8,813 question–answer pairs. Checkpoint selection uses both validation answer loss and the evaluation results for auxiliary fact predictions.

B.4 EM-BENCH INTERFACE

For the EM-Bench comparison, we adopt the same direct encoder-to-LLM interface as MERLIN. The EM encoder converts the IQ recording into signal tokens, which are projected into the LLM embedding space and passed directly to Qwen3.5-4B together with the question tokens. This configuration uses a per-token projector rather than the Q16 or Q128 query-based interaction module.

We train and validate our interface on the EM-134K corpus, which is split into a training set of 120,829 question–answer pairs from 95,679 signals and a validation set of 13,278 pairs from 10,587 signals. EM-Bench is held out as an independent test set; recordings in both EM-134K and EM-Bench use 512 or 1,024 IQ samples. These lengths yield 32 or 64 time-ordered patch features, respectively. A shared per-token projector maps each feature into the Qwen3.5-4B embedding space:

$$\mathrm{LN}(768) \to \mathrm{Linear}(768, 1536) \to \mathrm{GELU} \to \mathrm{Linear}(1536, 2560) \to \mathrm{LN}(2560).$$

All projected features are passed directly to Qwen3.5-4B.

Training comprises five stages. The first trains the projector with the EM Encoder and Qwen3.5-4B frozen. The next three stages progressively unfreeze the final two, final six, and all twelve EM Encoder blocks. The final stage jointly trains the EM Encoder, projector, and rank-16 Qwen3.5-4B LoRA. The Qwen3.5-4B base parameters remain frozen. Training samples are balanced across the EM-134K question families.

The projector peak learning rates across the five stages are 2×10^{-4}, 5×10^{-5}, 4×10^{-5}, 3×10^{-5}, and 2×10^{-5}, respectively. The peak learning rates at the top of the EM Encoder in the final four stages are 10^{-5}, 8×10^{-6}, 5×10^{-6}, and 2×10^{-6}, with layer-wise learning-rate decay of 0.8. The final joint stage uses a Qwen3.5-4B LoRA learning rate of 2×10^{-5}. Training runs on three H100 GPUs with a 5% warmup and gradient clipping at norm 1.0. Checkpoint selection gives equal weight to the ten-task perception group and the four-task generation group on the validation set.

C DATASETS AND EVALUATION PROTOCOLS

Table 9 summarizes the sizes, IQ lengths, and roles of the four datasets.

Table 9: Signal and QA counts used in the reported experiments.

Dataset	Train signals / QA	Validation signals / QA	Test signals / QA	IQ length	Primary use
LEMMO-ShortQA	3,832 / 15,328	870 / 3,480	846 / 3,384	128-5M	Q16 Recognition training and testing
LEMMO-LongDesc	16,103 / 165,328	856 / 8,813	652 / 802	0.4-8M	Q128 Description training and testing
EM-134K	95,679 / 120,829	10,587 / 13,278	–	512 or 1,024	EM-Bench training and validating
EM-Bench	–	–	3,370 / 4,200	512 or 1,024	EM-Bench testing

C.1 LEMMO-SHORTQA

LEMMO-ShortQA requires models to generate short answers from IQ signals and natural-language questions. It covers AMC, JPR, UAV recognition, and WTC, with 45 target labels in total. Table 10 reports the signal and QA counts for each task and split.

Table 10: LEMMO-ShortQA composition. Each physical signal has four question phrasings for the same task.

Task	Labels	Train signals / QA	Validation signals / QA	Test signals / QA
Modulation recognition	13	1,856 / 7,424	456 / 1,824	432 / 1,728
Jamming-pattern recognition	12	1,088 / 4,352	182 / 728	182 / 728
UAV recognition	10	120 / 480	40 / 160	40 / 160
Wireless-technology recognition	10	768 / 3,072	192 / 768	192 / 768
Total	45	3,832 / 15,328	870 / 3,480	846 / 3,384

Each physical signal has four fixed question phrasings for its task. Consequently, the QA count is four times the signal count in every split. For example, a modulation signal may be paired with questions such as "What modulation scheme does this signal use?" and "Identify the modulation class."

The modulation labels are 16PAM, 16QAM, 2FSK, 4PAM, AM-DSB, AM-DSB-SC, AM-LSB, AM-USB, CPFSK, FM, GFSK, PM, and QPSK.

The jamming-pattern labels are AM_Jamming, Comb_Jamming, FM_Jamming, Multi_Tone, Narrowband, No_Interference, Protocol_Bluetooth, Protocol_WiFi, Pulse_DME, Single_Tone, Sweep_Chirp, and Wideband.

The UAV labels are DJI Avata 2, DJI FPV Combo, DJI Mavic 3 Pro, Devention DEVO, Fly Sky FS I6X, Herelink HX4, JR Propo XG14, SIYI MK15, WFLY ET10, and Yunzhuo H12.

The wireless-technology labels are 5G-NR, DVB-T, IEEE 802.11ah, IEEE 802.15.4 SUN-OFDM, IEEE 802.15.4g, LTE, LoRa SF12, LoRa SF7, Sigfox, and Wi-Fi.

Table 11 lists the source datasets and the numbers of signals assigned to the training, validation, and test splits.

The modulation recordings contain 128–1,024 samples, the jamming recordings contain 128–16,368 samples, the wireless-technology recordings contain 4,096 samples, and the UAV recordings contain five million samples.

During the evaluation, each model generates an answer freely. The correctness is determined using the target labels for the corresponding task. We report the accuracy for each task and the mean accuracy across the four tasks. Compared models receive the same questions and label space.

C.2 LEMMO-LongDesc

LEMMO-LongDesc addresses long IQ observations and complex EM scenes. It contains five signal families: 4G LTE, 5G NR, basic-modulation scenes, ISM-band scenes, and UAV scenes. The dataset provides targeted QA on specific EM facts and questions requesting descriptions of the complete scene in a long IQ recording. Each signal and all its questions remain in the same split. Table 12 reports the distribution of signals and questions across the five families.

The training set contains 19,778 detailed description QA pairs and 145,550 targeted QA pairs. The validation set contains 1,006 description QA pairs and 7,807 targeted QA pairs. The test set contains 802 description QA pairs from 652 distinct physical signals.

The 4G LTE corpus supports recognition of radio-frequency activity and LTE signals, together with descriptions of observed communication activity. It covers LTE activity and ambient-noise scenes.

Table 11: Source datasets and signal counts for LEMMO-ShortQA.

Task	Source dataset	Train	Validation	Test
Modulation recognition	HisarMod2019.1 (Tekbıyık et al., 2020)	704	168	144
Modulation recognition	RML2016.04C (O'Shea et al., 2016)	384	96	96
Modulation recognition	RML2016.10A (O'Shea & West, 2016)	384	96	96
Modulation recognition	RML2016.10B (O'Shea & West, 2016)	384	96	96
Jamming-pattern recognition	CRAWDAD OWL-InIT / OWL Interference Dataset (Schmidt et al., 2019)	128	32	32
Jamming-pattern recognition	EM-Infer-Comm (Luo et al., 2025)	576	144	144
Jamming-pattern recognition	GNSS Jamming Dataset (Swinney & Woods, 2021)	384	6	6
UAV recognition	RFUAV (Shi et al., 2025)	120	40	40
Wireless-technology recognition	Real-world Wireless Communication Dataset: IEEE 802.11ax, LTE, and 5G-NR Signals (Subray, 2023)	192	48	48
Wireless-technology recognition	Sub-GHz IQ signals dataset / Multi-band sub-GHz technology recognition dataset (Fontaine et al., 2020)	384	96	96
Wireless-technology recognition	TechRec / Technology Recognition Dataset (Fontaine et al., 2019; Liu et al., 2017)	192	48	48

Table 12: LEMMO-LongDesc composition. The test column reports long-signal detailed description QA.

Signal family	Train signals / all QA	Validation signals / all QA	Test signals / description QA
4G LTE	2,940 / 35,280	235 / 2,820	120 / 120
5G NR	3,675 / 54,144	150 / 2,225	150 / 150
Basic-modulation scenes	3,675 / 25,725	150 / 1,050	150 / 300
ISM-band scenes	3,675 / 33,075	150 / 1,350	150 / 150
UAV scenes	2,138 / 17,104	171 / 1,368	82 / 82
Total	16,103 / 165,328	856 / 8,813	652 / 802

The 5G NR corpus includes questions about NR-signal presence, frequency, bandwidth, uplink and downlink properties, duplexing, subcarrier spacing, radio frames, visible terminals or SRS, and resource occupancy.

A basic-modulation observation window may contain multiple component signals. Its questions address the number of active signals, time-ordered signal identifiers $S_1, \ldots, S_n$, modulation types, SNR levels, time–frequency locations, and overlap relationships. Some reference answers provide both natural-language descriptions and structured JSON.

The ISM-band corpus covers the number and types of Wi-Fi, Bluetooth, and related signal sources, as well as protocols, bandwidths, channels, SNRs, time–frequency overlap, and user counts. The UAV corpus uses DroneRFa recordings and covers UAV-related signal presence, device models, and scene descriptions.

The first four signal families are generated with MATLAB; the UAV corpus uses DroneRFa recordings. IQ recordings in LEMMO-LongDesc contain approximately 0.4–8 million complex samples.

Model inputs comprise the IQ recording, acquisition-setting text, and the user question. The acquisition-setting text specifies the receiver center frequency, sampling rate, observation bandwidth, and signal duration. Signal duration is computed from IQ length and sampling rate.

Evaluation of long-signal descriptions uses free-form generation and assesses the accuracy of specific EM facts in each response. Compared models receive identical acquisition-setting text, system instructions, and questions. LEMMO processes each complete IQ recording. Merlin-LoRA supports input sequences of up to 4,096 IQ samples and therefore processes the first 4,096 samples of each long recording.

C.3 EM-134K

EM-134K is a public EM signal–text instruction dataset. Its released files contain 134,107 QA pairs. Its training and validation splits support model training and validation-stage model selection. Table 13 gives the split statistics.

Table 13: EM-134K split statistics.

Split	Distinct IQ signals	QA pairs
Train	95,679	120,829
Validation	10,587	13,278
Total	106,266	134,107

Each record includes fields such as `id`, `human_prompt`, `gpt_response`, `samples`, `answer`, `sample_rate`, and `snr`. The `samples` field stores the IQ recording, `human_prompt` contains the question, and `gpt_response` contains the textual response. The `answer` field provides the answer key for multiple-choice questions. IQ recordings contain 512 or 1,024 samples. The dataset covers option-based recognition, parameter estimation, interference analysis, and strategy generation.

Table 14 reports the number of QA pairs in each content category.

EM-134K includes public signals, professional simulations, and real-world recordings.

C.4 EM-BENCH

EM-Bench is a public test benchmark with 14 subtasks and 300 questions per subtask, yielding 4,200 test questions. Physical signals may be reused across subtasks; the benchmark contains 3,370 distinct IQ recordings. We use EM-Bench for testing. Table 15 lists the ten perception subtasks and four reasoning subtasks.

The ten perception subtasks use five-way multiple-choice questions, typically with four specific candidate answers and one "unable to answer" option. We evaluate these subtasks using option

Table 14: EM-134K QA categories.

Category	QA pairs	Representative objective
Modulation recognition (MOD)	9,996	Identify the modulation type
Bandwidth (PE_bandwidth)	1,250	Estimate occupied bandwidth
Duty cycle (PE_duty_cycle)	1,250	Estimate duty cycle
Number of pulses (PE_num_pulses)	6,000	Count radar pulses
Pulse repetition frequency (PE_prf)	8,500	Estimate PRF
Pulse width (PE_pulse_width)	12,000	Estimate pulse width
Signal-to-noise ratio (PE_snr)	7,250	Estimate SNR
First-pulse delay (PE_time_delay)	6,000	Estimate time delay
Protocol recognition (PI_protocol)	9,527	Identify a protocol or standard
Fragment detection (SD_fragment_detection)	9,023	Locate interference-segment boundaries
Joint modulation and interference analysis	6,250	Identify two signal attributes
Interference recognition	24,437	Identify interference presence and type
Other	986	Assess other signal attributes
Interference and countermeasure reasoning	31,638	Identify interference and generate countermeasures

Table 15: EM-Bench subtasks.

Capability	Subtask	Questions	Task requirement
Perception	MOD	300	Modulation recognition
Perception	PE_Radar_Bandwidth	300	Radar bandwidth estimation
Perception	PE_Radar_Duty_Cycle	300	Duty-cycle estimation
Perception	PE_Radar_Number_of_Pulses	300	Pulse-count estimation
Perception	PE_Radar_PRF	300	PRF estimation
Perception	PE_Radar_Pulse_Width	300	Pulse-width estimation
Perception	PI	300	Protocol or signal-standard recognition
Perception	CJR	300	Communication-jamming recognition
Perception	RJR	300	Radar-jamming recognition
Perception	SD	300	Interference-segment detection
Reasoning	Anti_CJ	300	Generate communication anti-jamming strategies
Reasoning	Anti_RJ	300	Generate radar anti-jamming strategies
Reasoning	CJS	300	Analyze and generate communication-jamming strategies
Reasoning	RJS	300	Analyze and generate radar-jamming strategies

accuracy. The four reasoning subtasks require text-based strategy generation and use ROUGE-L F1. All IQ recordings contain 512 or 1,024 samples.

D ADDITIONAL RESULTS OF LEMMO-LONGDESC

Additional results of LEMMO-LongDesc are shown below.

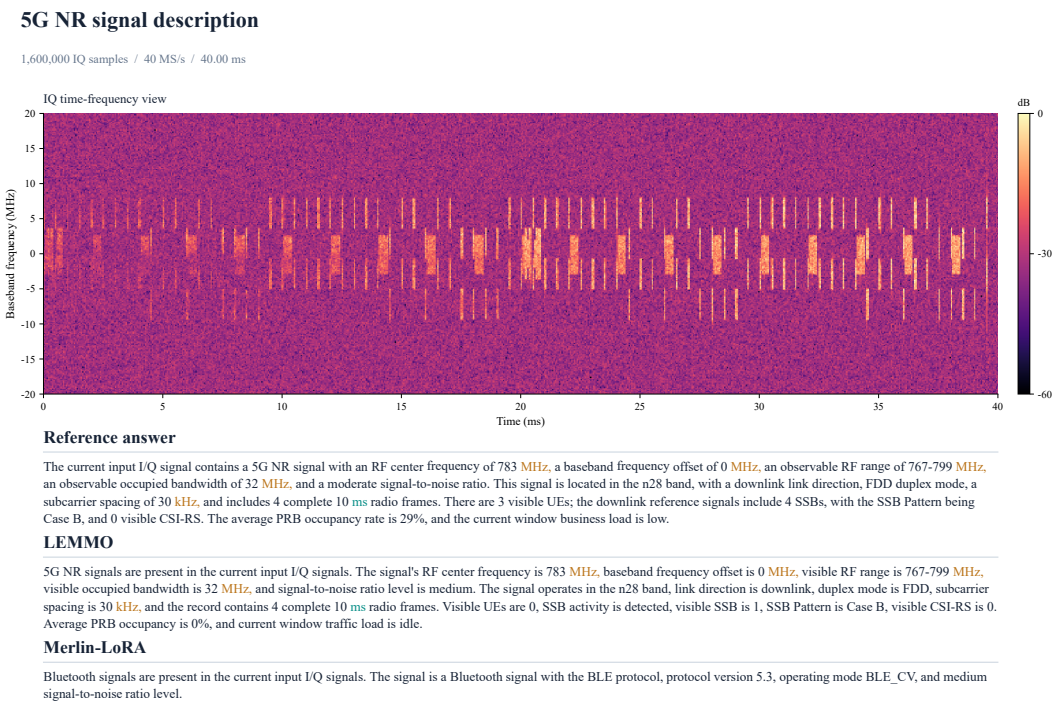
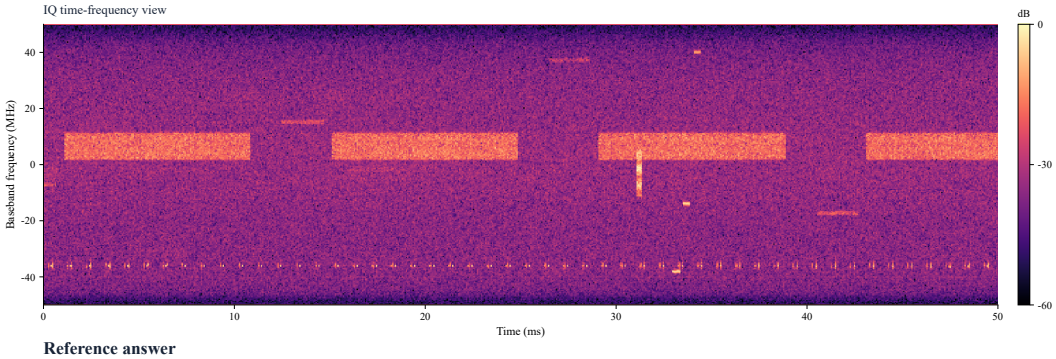


Figure 2: Case 1.

UAV signal description

5,000,000 IQ samples / 100 MS/s / 50.00 ms



Reference answer

RF signals related to the UAV were detected in the input I/Q signals, containing the FCS flight control link and the VTS video transmission link. The identified proprietary communication protocol is Lightbridge, with the possible UAV model being DJI Inspire 2. The FCS flight control link comprises a total of 3 frequency-hopping segments, distributed within the time range of 12.49-42.67 ms, with a center frequency coverage of 2422.70-2477.07 MHz, an occupied bandwidth of approximately 1.29-1.33 MHz, and individual frequency-hopping signal durations of approximately 2.18-2.22 ms. At the system level, the FCS employs FHSS spectrum access, while individual frequency-hopping segments utilize OFDM modulation. The VTS video transmission link employs OFDM modulation, with a center frequency of 2446.40 MHz and an occupied bandwidth of approximately 10 MHz.

LEMMO

RF signals related to the UAV were detected in the input I/Q signals, containing the FCS flight control link and the VTS video transmission link. The identified proprietary communication protocol is Lightbridge, with the possible UAV model being DJI Inspire 2. The FCS flight control link comprises a total of 4 frequency-hopping segments, distributed within 1.03-45.19 ms, with a center frequency coverage of 2414.19-2468.59 MHz, an occupied bandwidth of approximately 1.28-1.34 MHz, and individual frequency-hopping signal durations of approximately 2.18-2.20 ms. At the system level, the FCS employs FHSS spectrum access, with a single frequency-hopping segment using OFDM modulation. The VTS video transmission link employs OFDM modulation, with a center frequency of 2446.49 MHz and an occupied bandwidth of approximately 10 MHz.

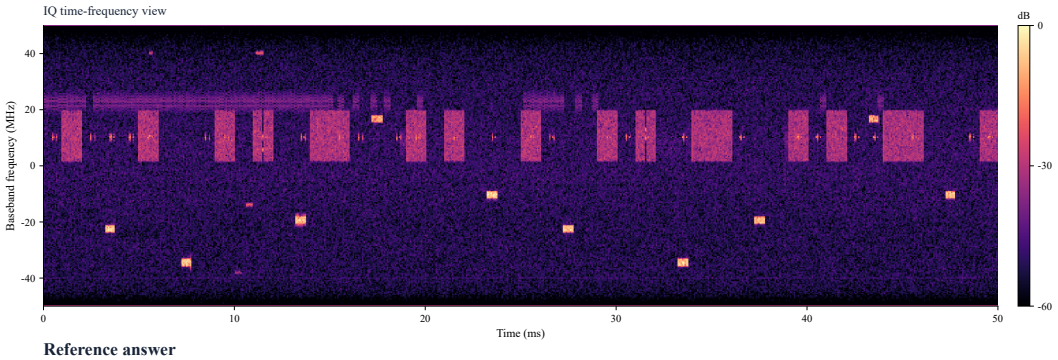
Merlin-LoRA

RF signals related to a drone were detected in the input I/Q signals, including an FCS flight control link, while no VTS video transmission link was detected. The possible drone model is FrSky X20. The FCS flight control link comprises a total of 7 frequency-hopping segments, distributed within a time range of 0.00-44.41 ms, with a center frequency coverage of 2405.39-2471.96 MHz, an occupied bandwidth of approximately 0.62-0.81 MHz, and individual frequency-hopping signal durations of approximately 1.86-1.90 ms. At the system level, the FCS employs FHSS for spectrum access, while individual frequency-hopping segments utilize CSS modulation.

Figure 3: Case 2.

UAV signal description

5,000,000 IQ samples / 100 MS/s / 50.00 ms



Reference answer

RF signals related to the UAV were detected in the input I/Q signals, containing both the FCS flight control link and the VTS video transmission link. The identified proprietary communication protocol is OcuSync 3.0, with the possible UAV model being DJI Mini 3 Pro. The FCS flight control link comprises a total of 10 frequency-hopping segments, distributed within a time range of 3.20-47.76 ms, with a center frequency coverage of 2405.48-2456.45 MHz, an occupied bandwidth of approximately 2.57-2.75 MHz, and individual frequency-hopping signal durations of approximately 0.58-0.64 ms. At the system level, the FCS employs FHSS spectrum access, while individual frequency-hopping segments utilize OFDM modulation. The VTS video transmission link employs OFDM modulation, with a center frequency of 2450.50 MHz and an occupied bandwidth of approximately 20 MHz.

LEMMO

RF signals related to the UAV were detected in the input I/Q signals, containing both the FCS flight control link and the VTS video transmission link. The identified proprietary communication protocol is OcuSync 3.0, with the possible UAV model being DJI Mini 3 Pro. The FCS flight control link comprises a total of 10 frequency-hopping segments, distributed within a time range of 1.03–47.64 ms, with a center frequency coverage of 2405.48–2450.50 MHz, an occupied bandwidth of approximately 2.52–2.70 MHz, and individual frequency-hopping signal durations of approximately 0.58–0.64 ms. At the system level, the FCS employs FHSS spectrum access, while individual frequency-hopping segments utilize OFDM modulation. The VTS video transmission link employs OFDM modulation, with a center frequency of 2450.69 MHz and an occupied bandwidth of approximately 20 MHz.

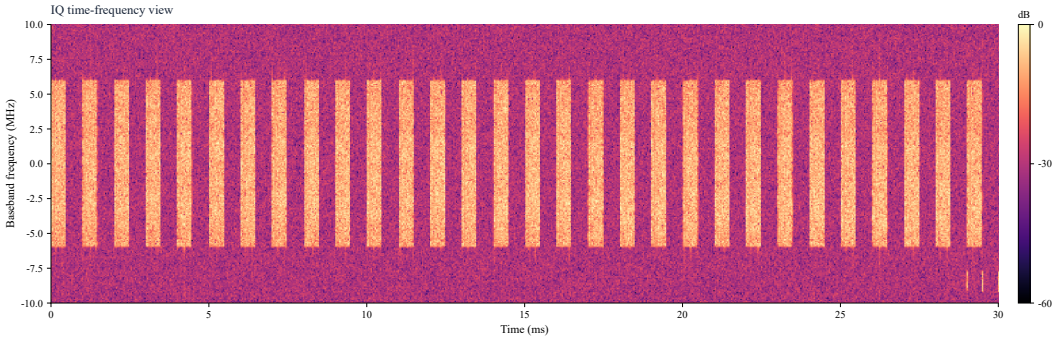
Merlin-LoRA

RF signals related to the UAV are detected in the input I/Q signals, containing the FCS flight control link and the VTS video transmission link. The identified proprietary communication protocol is Lightbridge, and the possible UAV model is DJI MATRICE 200. The FCS flight control link comprises a total of 3 frequency-hopping segments, distributed within the time range of 11.30-40.84 ms, with a center frequency coverage of 2405.47-2463.45 MHz, an occupied bandwidth of approximately 2.42-2.56 MHz, and individual frequency-hopping signal durations of approximately 2.18-2.20 ms. At the system level, the FCS employs FHSS spectrum access, and each frequency-hopping segment utilizes OFDM modulation. The VTS video transmission link employs OFDM modulation, with a center frequency of 2450.5 MHz and an occupied bandwidth of approximately 10 MHz.

Figure 4: Case 3.

5G NR signal description

600,000 IQ samples / 20 MS/s / 30.00 ms



Reference answer

A 5G NR signal is detected in the current input I/Q signal. Its RF center frequency is 3350 MHz, baseband frequency offset is 0 MHz, visible RF range is 3340-3360 MHz, visible occupied bandwidth is 20 MHz, and signal-to-noise ratio level is strong. The signal operates in the n77/n78 band, link direction is uplink, duplex mode is TDD, subcarrier spacing is 30 kHz, and the record contains 3 complete 10 ms radio frames. Three visible UEs can be confirmed, including 3 visible SRS. Average PRB occupancy is 65%, and current window traffic load is medium.

LEMMO

A 5G NR signal is detected in the current input I/Q signals. The signal's RF center frequency is 3350 MHz, baseband frequency offset is 0 MHz, visible RF range is 3340-3360 MHz, visible occupied bandwidth is 20 MHz, and SNR level is strong. Its NR frequency band is n77/n78, link direction is uplink, duplex mode is TDD, SCS is 30 kHz, and there are 3 complete 10 ms radio frames. Visible UEs are 3, visible SRS are 3. Average PRB occupancy rate is 65%, and current window business load is medium.

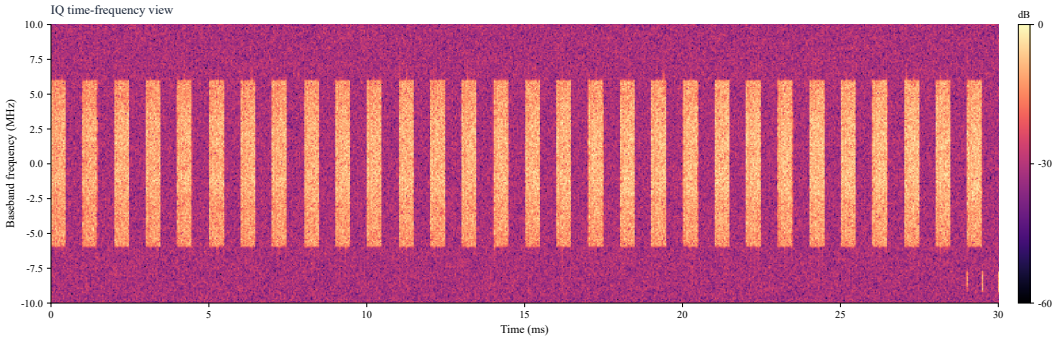
Merlin-LoRA

The input I/Q signal contains a 5G NR signal. The signal RF center frequency is 3350 MHz, the baseband frequency offset is 0 MHz, the visible RF range is 3340-3360 MHz, the visible occupied bandwidth is 20 MHz, and the SNR level is medium. The signal operates in the n77/n78 bands, the link direction is uplink, the duplex mode is TDD, the SCS is 30 kHz, and the record contains 3 complete 10 ms radio frames. The visible UE is 1, the visible SRS is 1. The average PRB occupancy rate is 65%, and the current window business load is medium.

Figure 5: Case 4.

5G NR signal description

600,000 IQ samples / 20 MS/s / 30.00 ms



Reference answer

The input I/Q signal contains a 5G NR signal with an RF center frequency of 3350 MHz, a baseband frequency offset of 0 MHz, an observable RF range of 3340-3360 MHz, an observable occupied bandwidth of 20 MHz, and a strong signal-to-noise ratio. The signal operates in the n77/n78 bands, with the link direction being uplink, TDD duplex mode, and a subcarrier spacing of 30 kHz. The recorded data includes three complete 10 ms radio frames. Within the current window, three UEs and three SRS are visible. Simultaneously, the average PRB occupancy is 65%, and the current window traffic load is moderate.

LEMMO

A 5G NR signal is detected in the current input I/Q signals. The signal's RF center frequency is 3350 MHz, baseband frequency offset is 0 MHz, visible RF range is 3340-3360 MHz, visible occupied bandwidth is 20 MHz, and SNR level is strong. Its NR frequency band is n77/n78, link direction is uplink, duplex mode is TDD, SCS is 30 kHz, and there are 3 complete 10 ms radio frames. Visible UEs are 3, visible SRS are 3. Average PRB occupancy rate is 65%, and current window business load is medium.

Merlin-LoRA

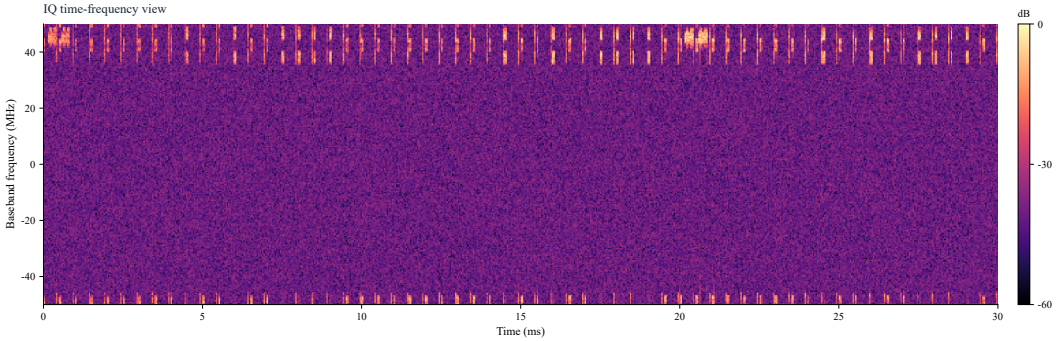
The input I/Q signal contains a 5G NR signal. The signal RF center frequency is 3350 MHz, the baseband frequency offset is 0 MHz, the visible RF range is 3340-3360 MHz, the visible occupied bandwidth is 20 MHz, and the signal-to-noise ratio level is weak. The signal operates in the n77/n78 bands, the link direction is uplink, the duplex mode is TDD, the SCS is 30 kHz, and the record contains 3 complete 10 ms radio frames. The visible UE is 1, the visible SRS is 1. The average PRB occupancy rate is 65%, and the current window business load is medium.

Figure 6: Case 5.

]

5G NR signal description

3,000,000 IQ samples / 100 MS/s / 30.00 ms



Reference answer

A 5G NR signal is detected in the current input I/Q signal, with an RF center frequency of 881 MHz, a baseband frequency offset of -55 MHz, an observable RF range of 886-891 MHz, an observable occupied bandwidth of 5 MHz, and a strong signal-to-noise ratio. The signal operates in the n5 band, with the link direction being downlink, FDD duplex mode, and a subcarrier spacing of 30 kHz. The record contains three complete 10 ms radio frames. Within the current window, zero visible UEs are present, while SSBs exist, including four visible SSBs; the SSB pattern is Case B, and one visible CSI-RS is present. The average PRB occupancy rate within the current window is 0%, and the traffic load is idle.

LEMMO

A 5G NR signal is present in the current input I/Q signals. The signal RF center frequency is 881 MHz, the baseband frequency offset is -55 MHz, the visible RF range is 886-891 MHz, the visible occupied bandwidth is 5 MHz, and the signal-to-noise ratio level is strong. This signal operates in the n5 band, the link direction is downlink, the duplex mode is FDD, the SCS is 30 kHz, and there are 3 complete 10 ms radio frames. The visible number of UEs is 0, SSB activity exists, the visible number of SSBs is 4, the SSB Pattern is Case B, and the visible number of CSI-RS is 0. The average PRB occupancy rate is 0%, and the current window business load is idle.

Merlin-LoRA

The input I/Q signal contains a 5G NR signal. The signal RF center frequency is 881 MHz, the baseband frequency offset is -55 MHz, the visible RF range is 886-891 MHz, the visible occupied bandwidth is 5 MHz, and the SNR level is strong. The signal operates in the n5 band, the link direction is downlink, the duplex mode is FDD, the SCS is 30 kHz, and there are 3 complete 10 ms radio frames. The visible UE count is 0, SSB activity is present, the visible SSB count is 1, the SSB Pattern is Case A, the visible CSI-RS count is 0. The average PRB occupancy rate is 0%, and the current window business load is idle.

Figure 7: Case 6.