A. Appendix

A.1. Overview of Appendix

This appendix provides supplementary details for our pre-trained baseline, dataset, and benchmark. Due to space constraints in the main paper, additional technical specifics can only be accommodated here.

The organization is as follows:

- Sec. A.2: More implementation details of baseline model.
- Sec. A.3: Datasheets for benchmark.
- Sec. A.4: SNR Robustness: Pretrain Model vs. MERLIN.
- Sec. A.5: Visualizations of samples.

A.2. Implementation Details of Baseline Model

We utilized EM-134K as the supervised fine-tuning (SFT) dataset. The pre-trained model is derived from Qwen3-4B-Instruct-2507, to which we first integrated a signal encoder (SIT) that embeds IQ signal data into feature vectors. These feature vectors serve as a second modality input alongside text into the LLM.

During two-stage training, we perform full fine-tuning in the first stage to align signal and text semantics. In the second stage, we apply knowledge distillation fine-tuning to enhance the model's robustness under low-SNR conditions, where the LLM parameters are frozen and only the signal encoder and projection layers are trained.

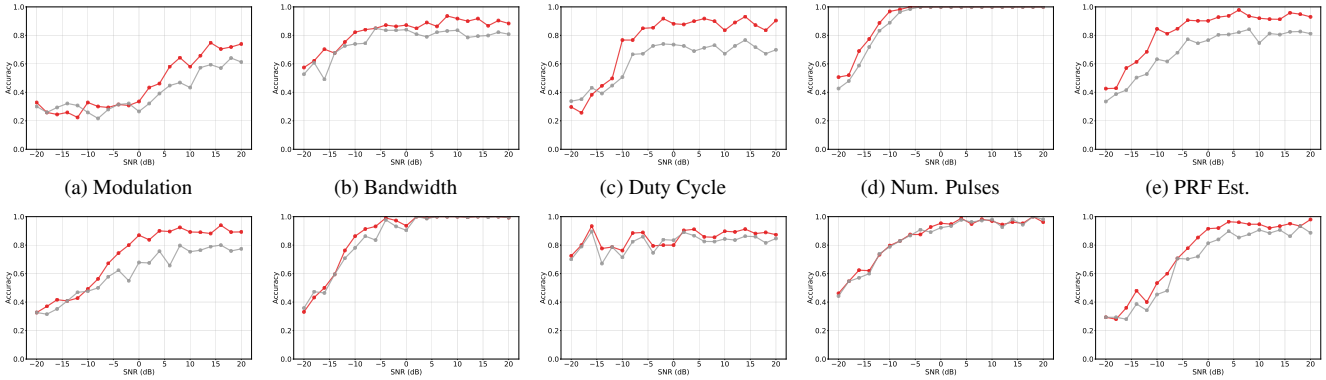
Table 4. **Training configuration details for MERLIN framework.** Hyperparameters used during the two-stage training. SIT = Signal Integration Transformer encoder.

Configuration	Parameter Value
Dataset (Stage 1)	EM-134K (134K instruction pairs)
Dataset (Stage 2)	Synthetic low-SNR variants (SNR = -20dB to 20dB)
Batch Size	256
LR: Signal Encoder (SIT)	5.0×10^{-5} (cosine decay)
LR: Projector, LLM	5.0×10^{-5}
Optimizer	AdamW ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon =$ 1e-8)
Training Epochs	Stage 1: 8 epochs, Stage 2: 8 epochs
Hardware	$8 \times$ NVIDIA A100 80GB

A.3. Datasheets for benchmark

Table 5. **Overview of the EM-Bench evaluation framework.** The benchmark features a three-level hierarchy spanning Perception and Reasoning, consisting of 14 distinct L3-Tasks. Each sub-task contains 300 samples, totaling 4,200 VQA pairs. The table lists the specific abbreviations, data quantities, and answer types (Single Choice vs. Open-ended Strategy) for each task.

L1-Task	L2-Task	L3-Task	Abbr.	Annotation Format	Number of Samples	Answer Type
Perception	Signal Characterization	Modulation Classification	MOD	VQA	300	Single Choice(A/B/C/D/E)
		Duty Cycle Estimation	PE.DC	VQA	300	Single Choice(A/B/C/D/E)
		Pulse Repetition Frequency Estimation	PE.PRF	VQA	300	Single Choice(A/B/C/D/E)
		Bandwidth Estimation	PE.BW	VQA	300	Single Choice(A/B/C/D/E)
		Pulse Width Estimation	PE.PW	VQA	300	Single Choice(A/B/C/D/E)
		Pulse Number Estimation	PE.NoP	VQA	300	Single Choice(A/B/C/D/E)
		Protocol Identification	PI	VQA	300	Single Choice(A/B/C/D/E)
	Jamming Identification	Radar Jamming Judgement	RJR	VQA	300	Single Choice(A/B/C/D/E)
		Communication Jamming Judgement	CJR	VQA	300	Single Choice(A/B/C/D/E)
	Segment Detection	Jamming Segment Detection	SD	VQA	300	Single Choice(A/B/C/D/E)
Reasoning	Strategy Generation	Anti-Communication Jamming Strategy	Anti-CJ	VQA	300	Open-end Strategy
		Anti-Radar Jamming Strategy	Anti-RJ	VQA	300	Open-end Strategy
		Communication Jamming Strategy	CJS	VQA	300	Open-end Strategy
		Radar Jamming Strategy	RJS	VQA	300	Open-end Strategy



(a) Modulation (b) Bandwidth (c) Duty Cycle (d) Num. Pulses (e) PRF Est.

(f) Pulse Width (g) Protocol ID (h) Comm. Jamming (i) Radar Jamming (j) Segment Det.

Figure 6. **Detailed performance comparison across 10 sub-tasks.** The red line denotes MERLIN and the gray line denotes the Stage-1 baseline. MERLIN demonstrates significant improvements in complex parameter estimation tasks (e.g., Duty Cycle, Pulse Width) across all SNR levels, while maintaining robust performance in classification tasks.

A.4. Detailed Performance Analysis across SNR Levels

Figure 6 provides a comprehensive breakdown of model performance across 10 sub-tasks under varying SNR conditions. The comparison between MERLIN and the Stage-1 baseline reveals distinct behavioral patterns:

- **Superiority in Fine-grained Perception:** In complex parameter estimation tasks such as *Duty Cycle*, *Pulse Width*, and *PRF Estimation*, MERLIN exhibits a consistent and significant performance lead (often $> 10\%$) across the entire SNR spectrum. This indicates that our feature-level distillation not only improves noise robustness but also enhances the model's fundamental ability to extract subtle signal features that the baseline fails to capture even at high SNRs.
- **Robustness in Modulation and Detection:** For *Modulation Recognition (MOD)* and *Signal Detection (SD)*, MERLIN demonstrates stronger resilience. While both models degrade as noise increases, MERLIN maintains a clearer operational margin in challenging low-SNR environments (e.g., -10 dB to 0 dB).
- **Performance on Saturated Tasks:** For tasks like *Protocol Identification (PI)* and *Jamming Recognition (RJR)*, both models achieve high accuracy rapidly. However, MERLIN remains competitive and consistently matches or slightly exceeds the baseline, ensuring no degradation in simpler tasks while significantly boosting performance in complex ones.

A.5. Visualizations of samples

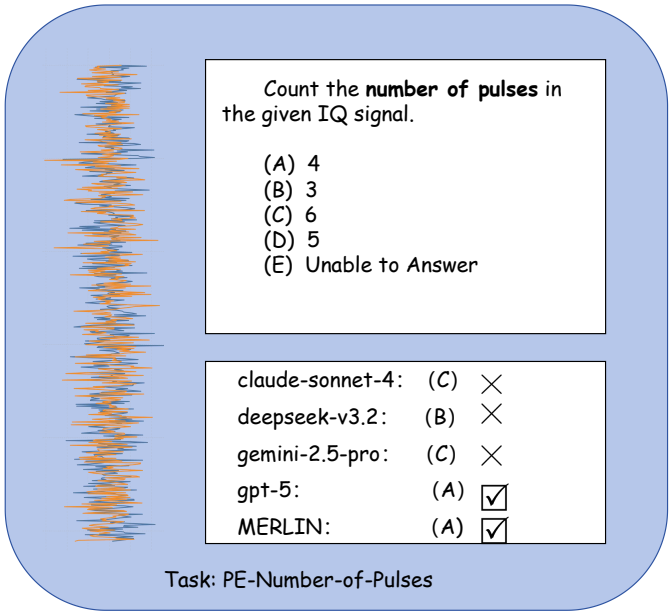


Figure 7. Number-of-Pulses estimation details

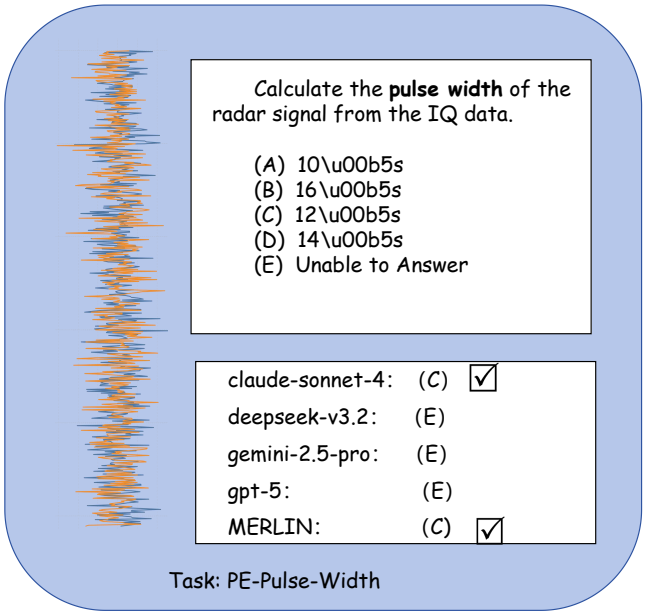


Figure 8. Pulse-Width estimation classification

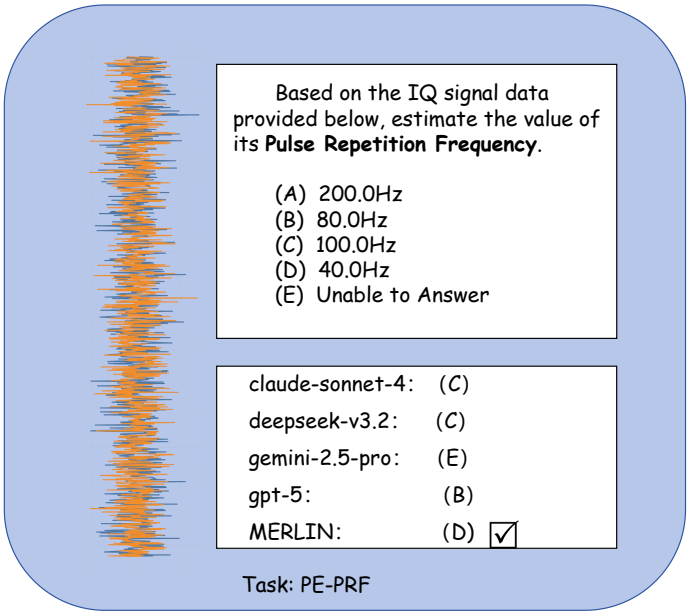


Figure 9. PRF estimation details

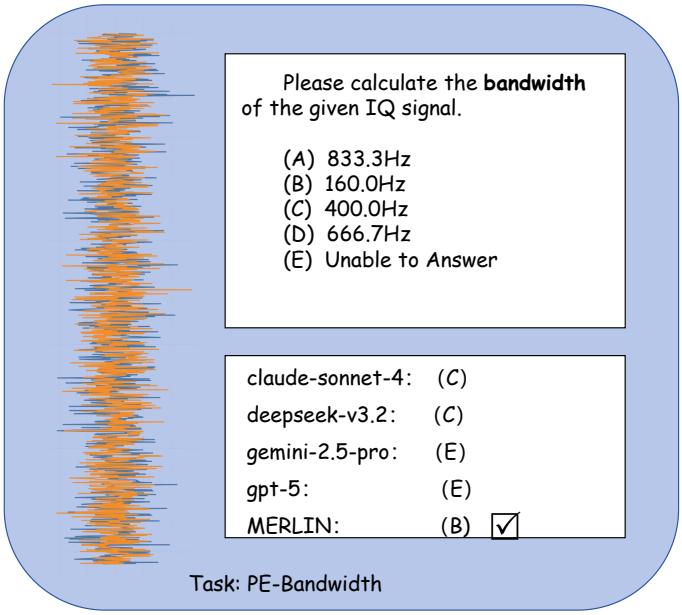


Figure 10. Bandwidth estimation details

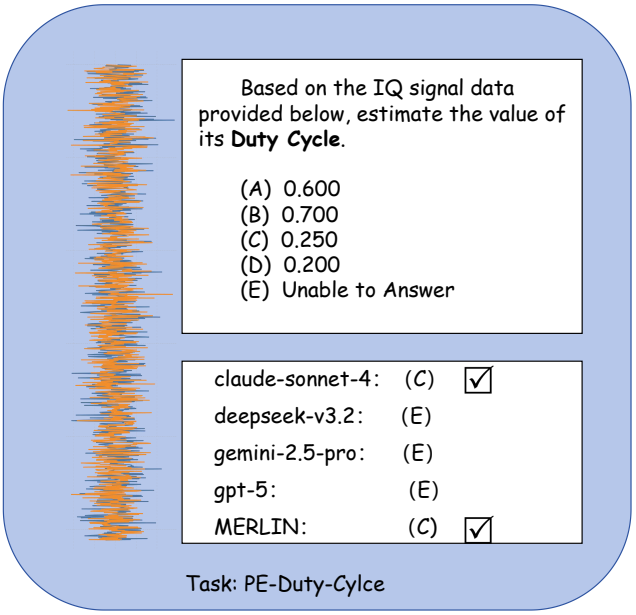


Figure 11. Duty-Cylce estimation details

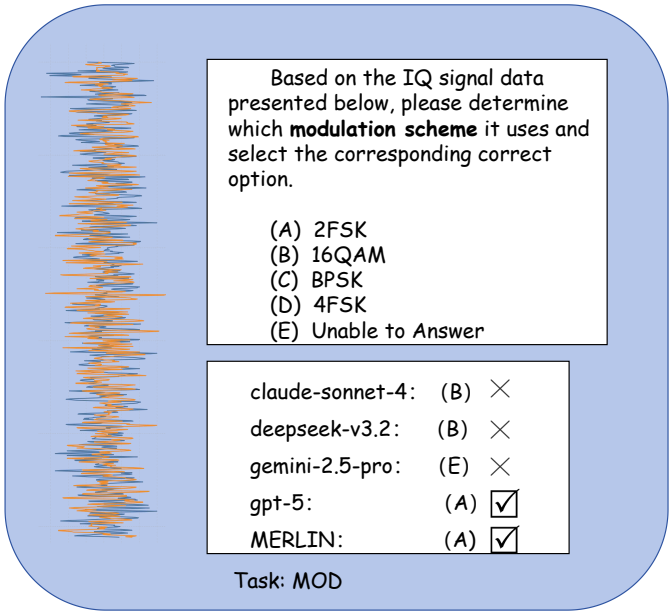


Figure 12. Modulation details

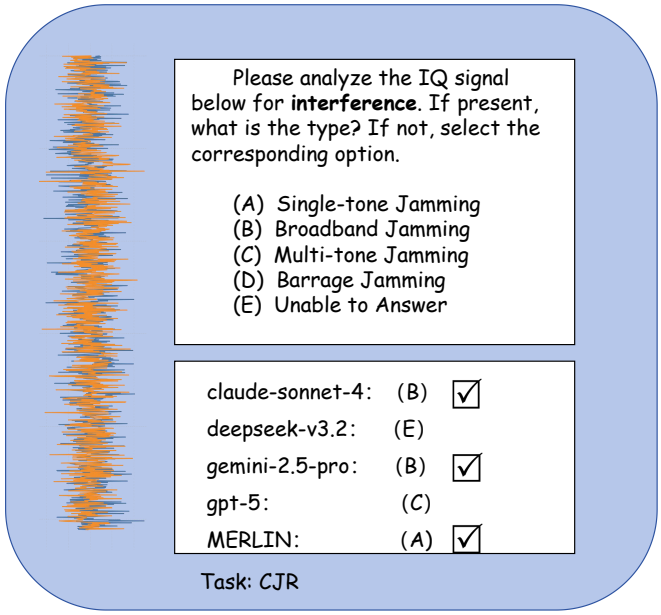


Figure 13. CJR details

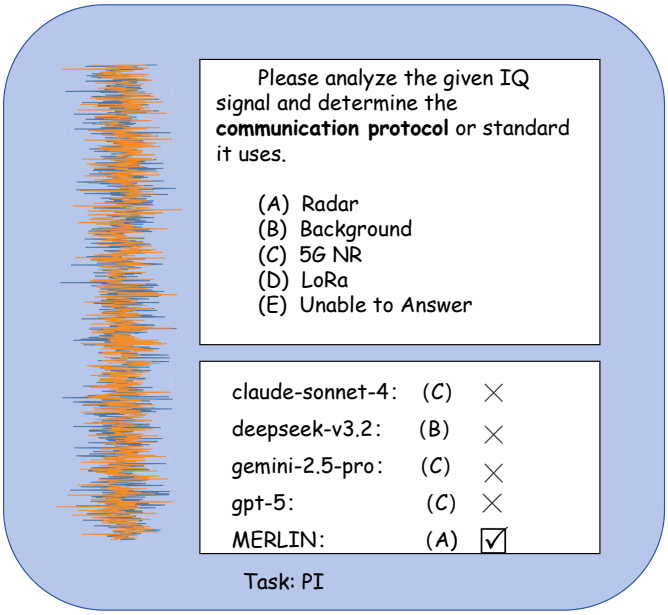


Figure 14. PI details

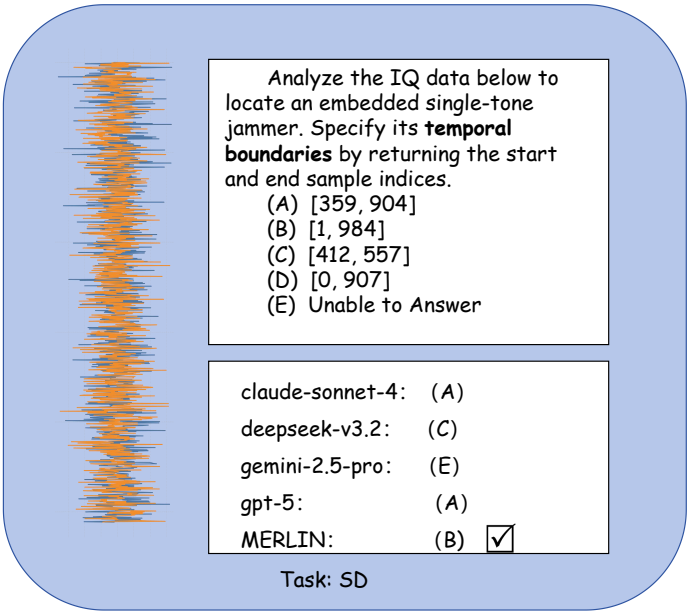


Figure 15. SD details

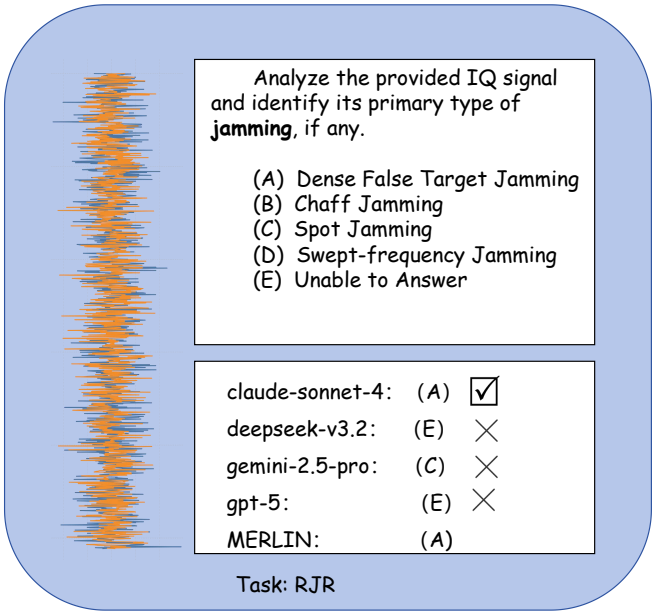


Figure 16. RJR details