Adapter Modules
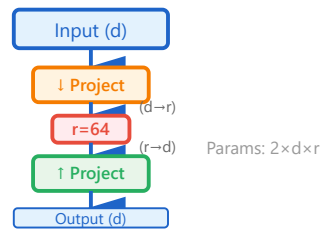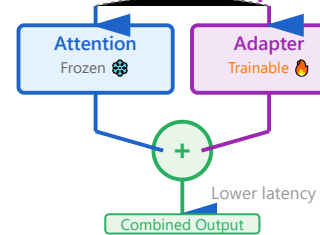
Insert Small Layers between frozen model layers
Parallel vs Sequential adapter architectures
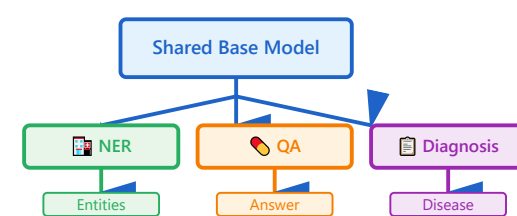
Sequential Adapter

Input (d)
↓ Project
(d→r)
r=64
(r→d)
Params: 2×d×r
↑ Project
Output (d)

Parallel Adapter

Attention
Frozen ❄️

Adapter
Trainable ✋

+

Lower latency

Combined Output

Multi-Task Adapters

Shared Base Model

📋 NER   💊 QA   📋 Diagnosis

Entities   Answer   Disease

1-5%
Added Params

64-512
Bottleneck Size

Modular
Architecture

🔧 Adapter Types

- **Sequential:** Down-project → ReLU → Up-project
- **Parallel:** Run alongside attention layers
- **Multi-task:** Separate adapters per task
- **Stacked:** Layer-wise adapter composition

🥼 Medical Task Adapters

- Clinical NER adapter
- Medical QA adapter
- Diagnosis prediction adapter
- Report generation adapter