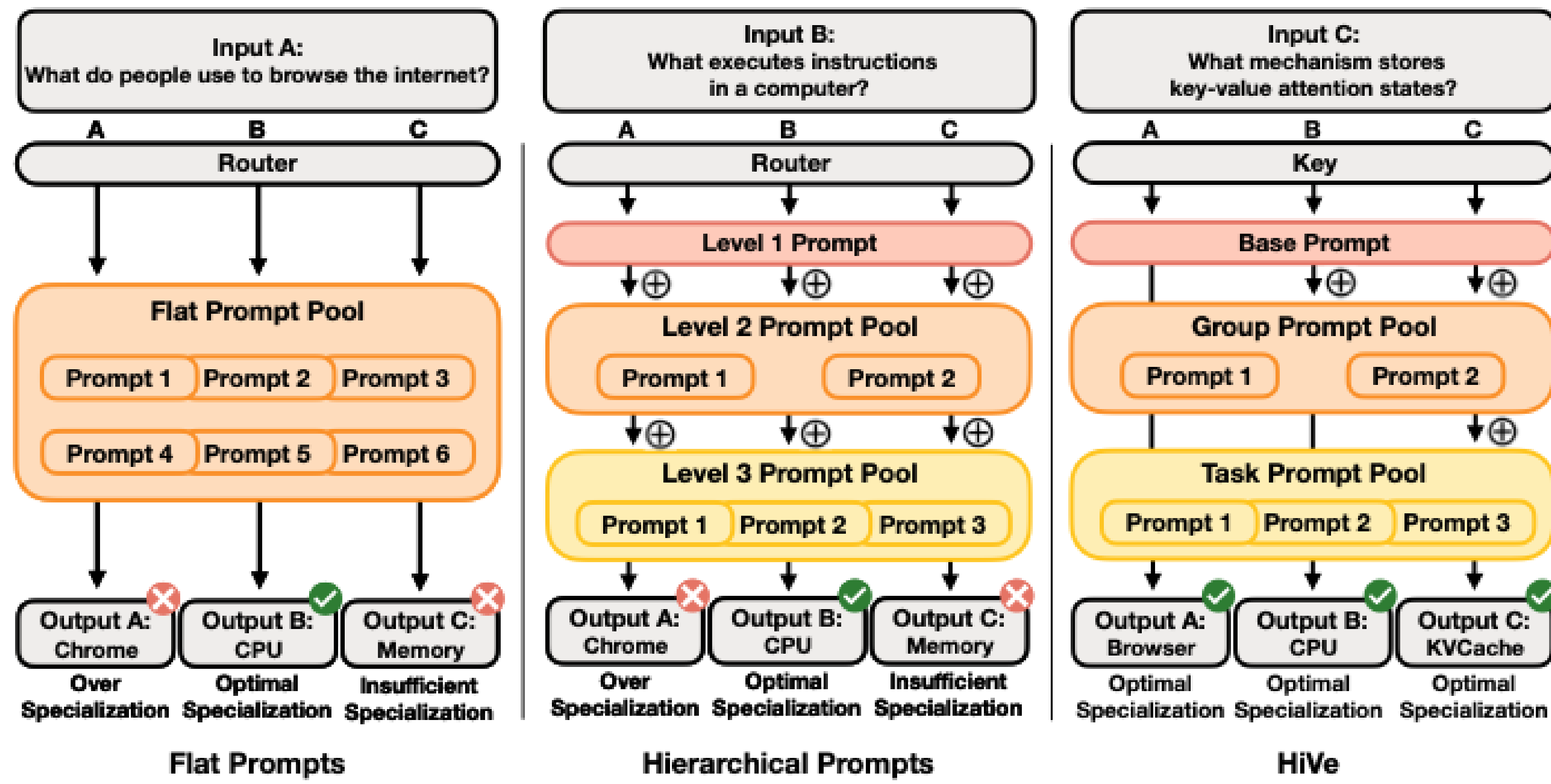




1. Motivation



Key Question

Why use the same prompt level for every input?

Flat Prompts

✗ Single level for all inputs

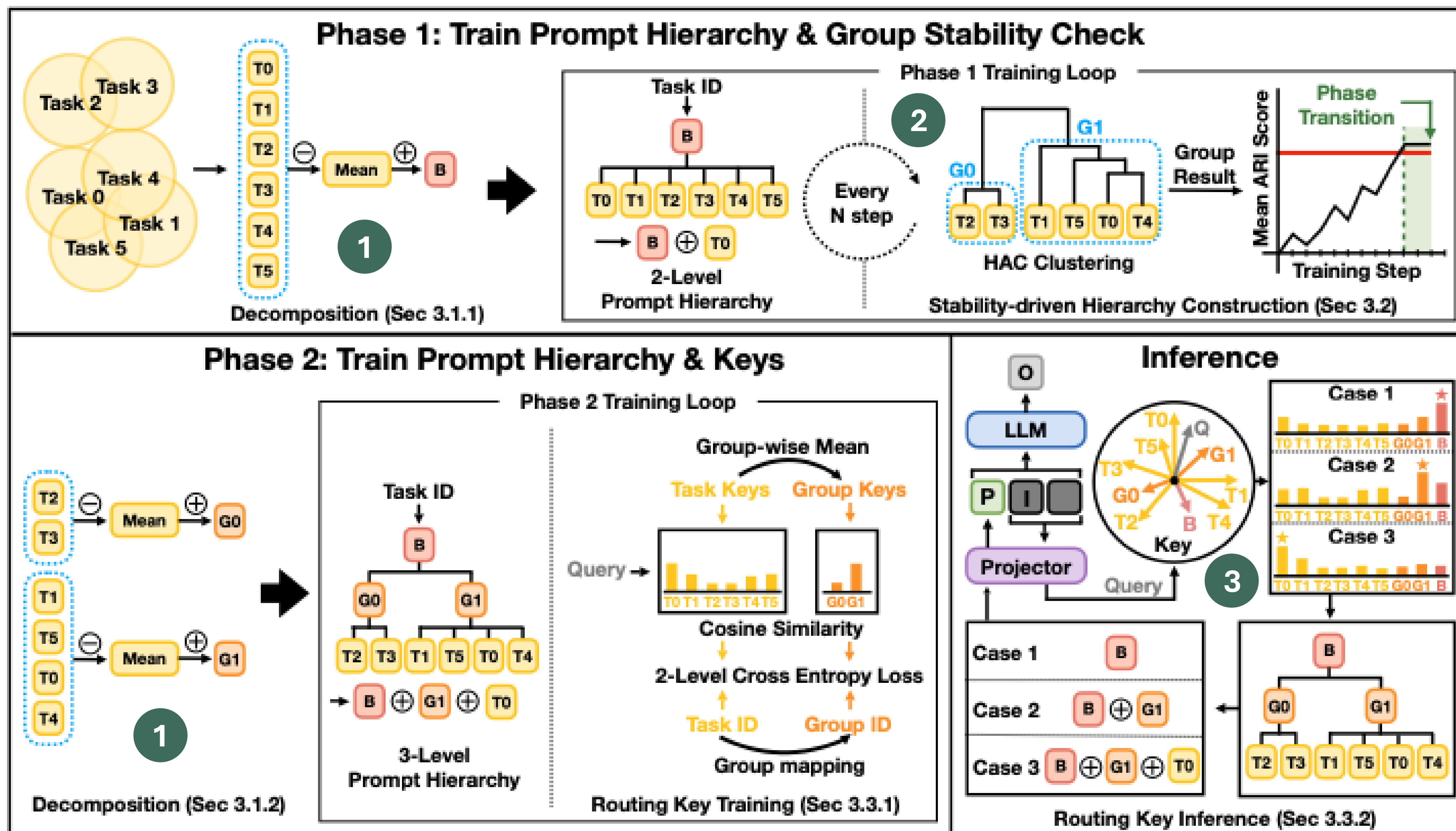
Hierarchical Prompts

✗ Fixed level for all inputs

HiVe

✓ Adaptive level per input

2. Methodology



① Residual Prompt Decomposition

- ⚠ Shared & task-specific knowledge are entangled
- 💡 Lift the shared mean to upper-level prompt
- 🎯 Less redundancy & interference

② Stability-driven Hierarchy Construction

- ⚠ Task relationships are unknown
- 💡 Group tasks once clustering is stable
- 🎯 Tasks with similar behavior grouped

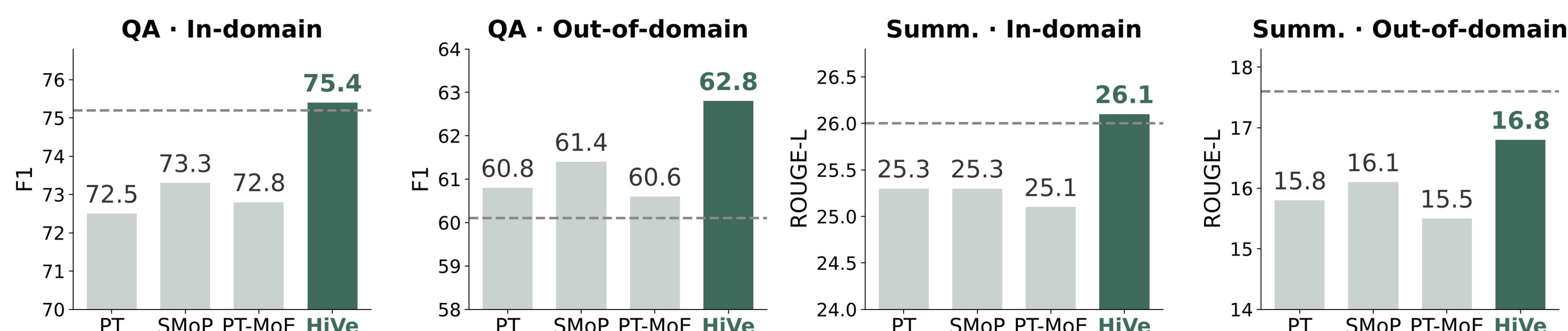
③ Vertical Mixture-of-Experts (V-MoE)

- ⚠ Inputs need different specialization
- 💡 Route by key: B / B+G / B+G+T
- 🎯 Input-adaptive prompt specialization

3. Experiments & Results

Main results vs. prompt-tuning baselines

----- Full fine-tuning (1.2B params)

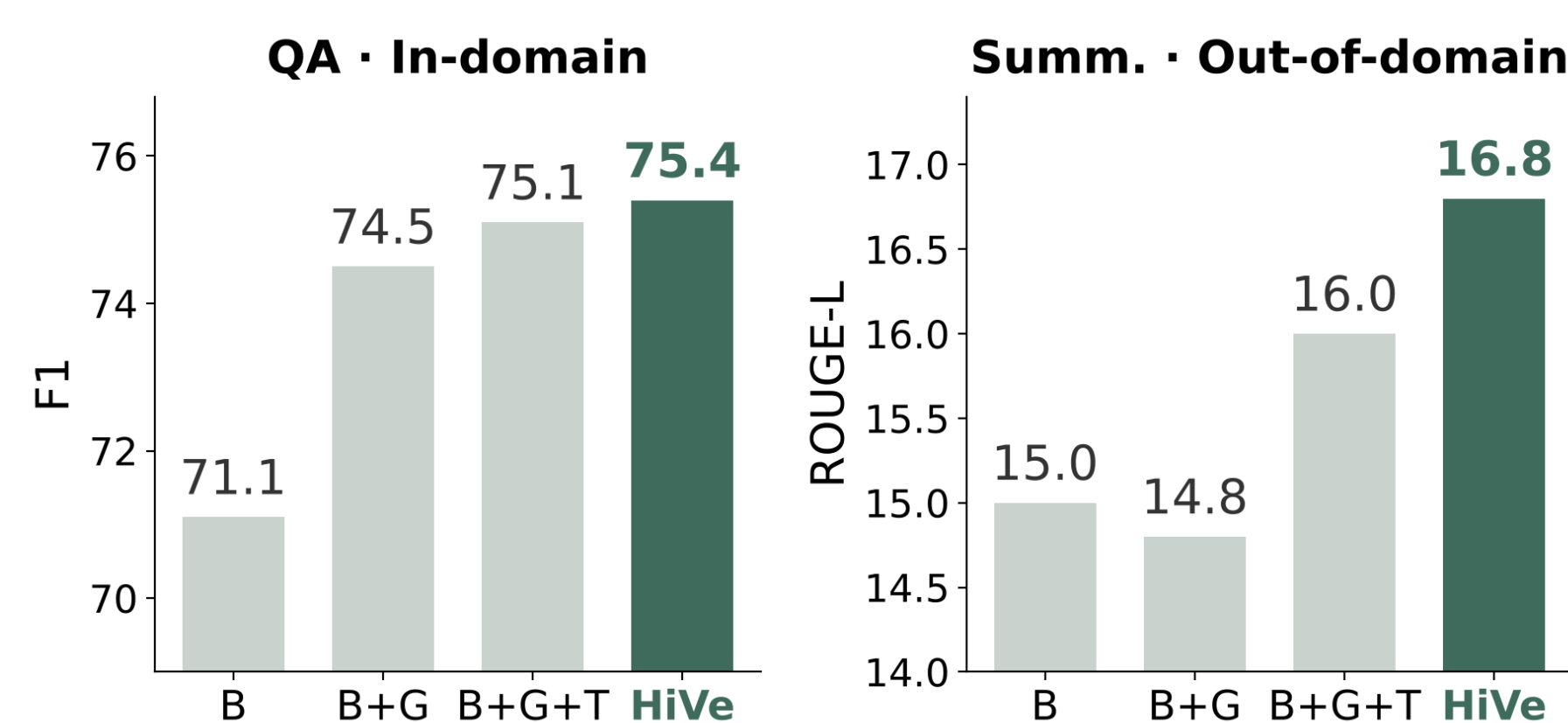
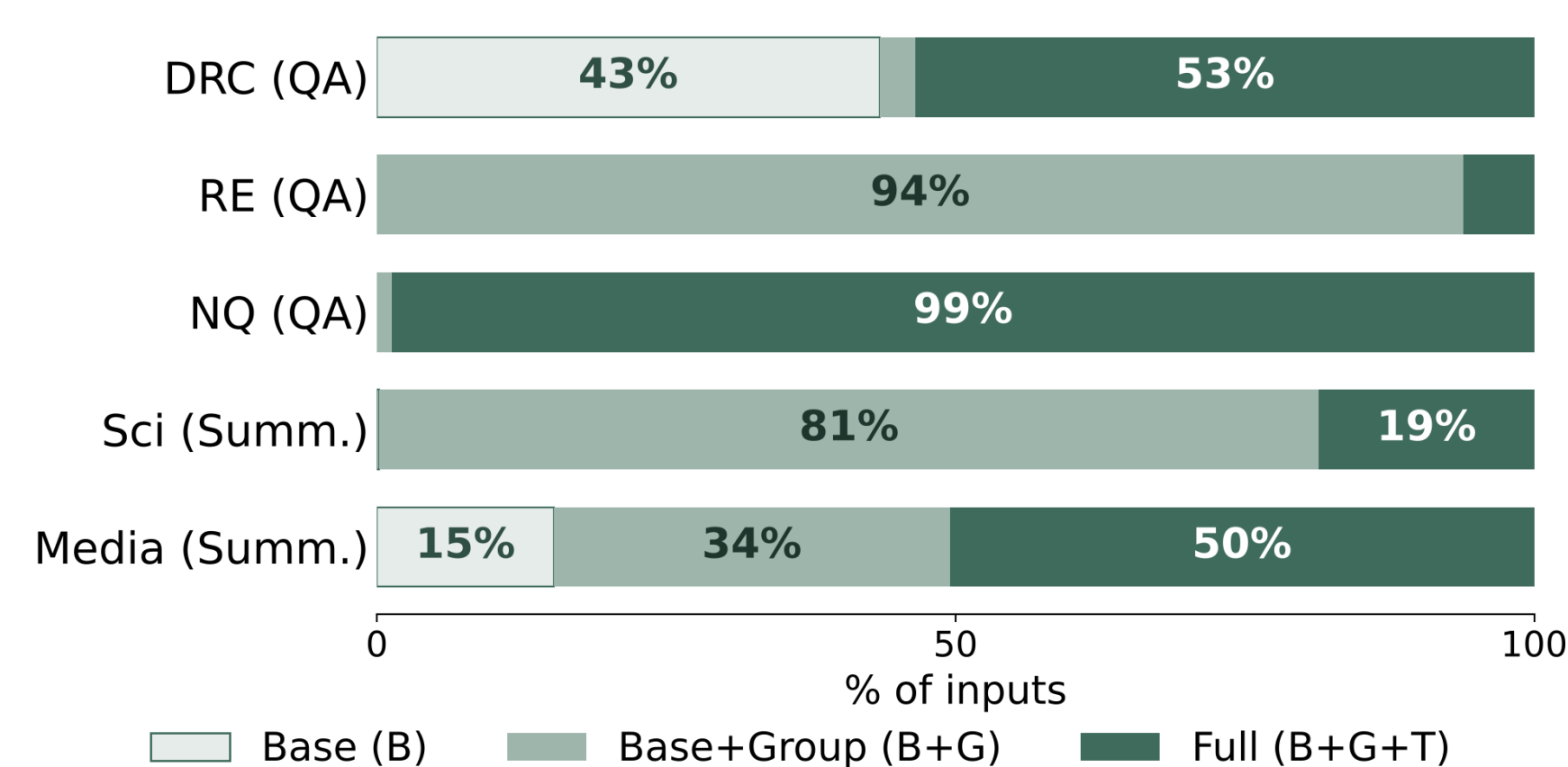


Main Results

- ✓ Best prompt-tuning results, in- & OOD
- ✓ Comparable to full FT with far fewer params

Which level does V-MoE pick?

Fixed level vs. adaptive (V-MoE)



V-MoE Analysis

- ✓ Diverse prompt levels across inputs
- ✓ Clear gains over fixed-level prompts