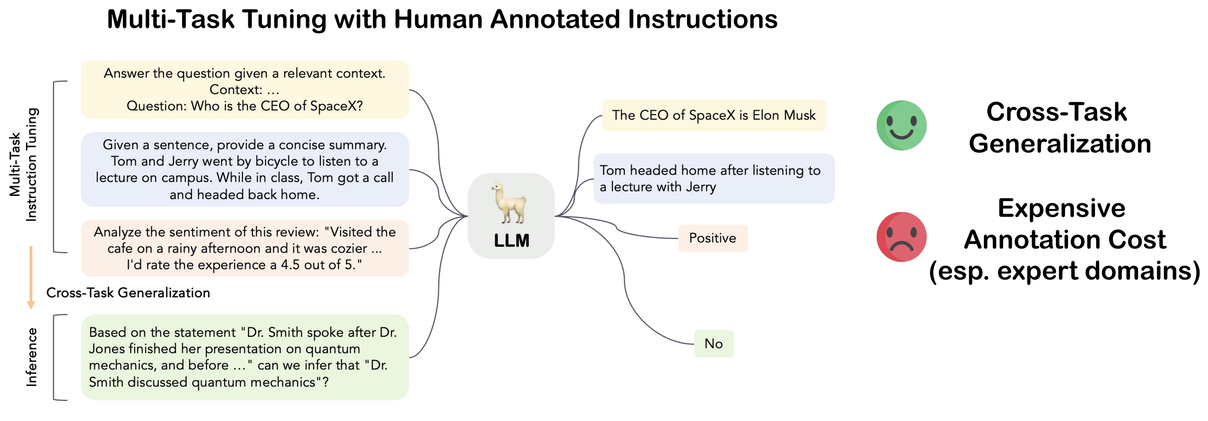
Self-Specialization: Uncovering Latent Expertise within Large Language Models

Junmo Kang[1] Hongyin Luo[2] Yada Zhu[3] Jacob Hansen[2] James Glass[2] David Cox[3]
Alan Ritter[1] Rogerio Feris[3] Leonid Karlinsky[3]

Junmo.kang@gatech.edu

Georgia Institute of Technology [1]
MIT CSAIL [2]
MIT-IBM Watson AI Lab [3]

Instruction-Tuning & Self-Alignment

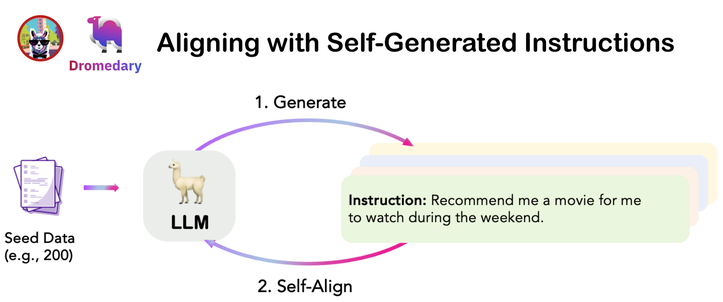


Cross-Task Generalization

Expensive Annotation Cost (esp. expert domains)

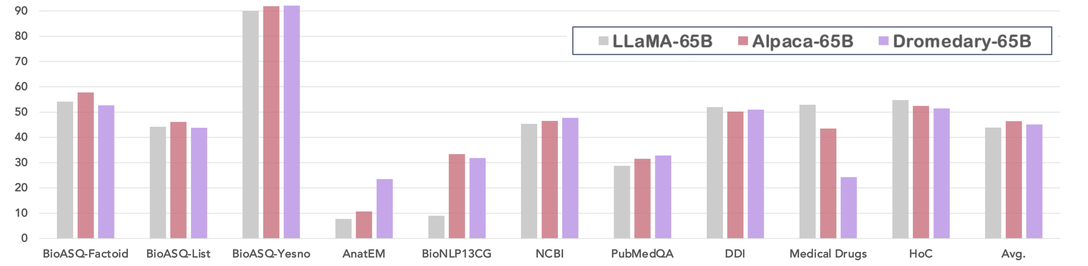
Minimal Human Supervision

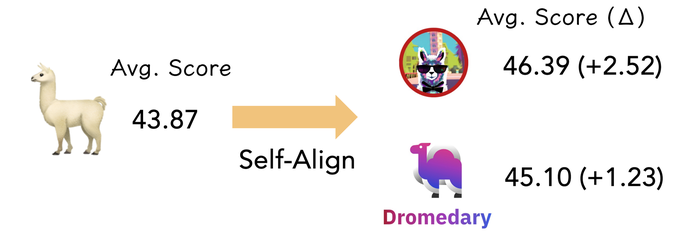
Created Instructions Too General



Preliminary Study

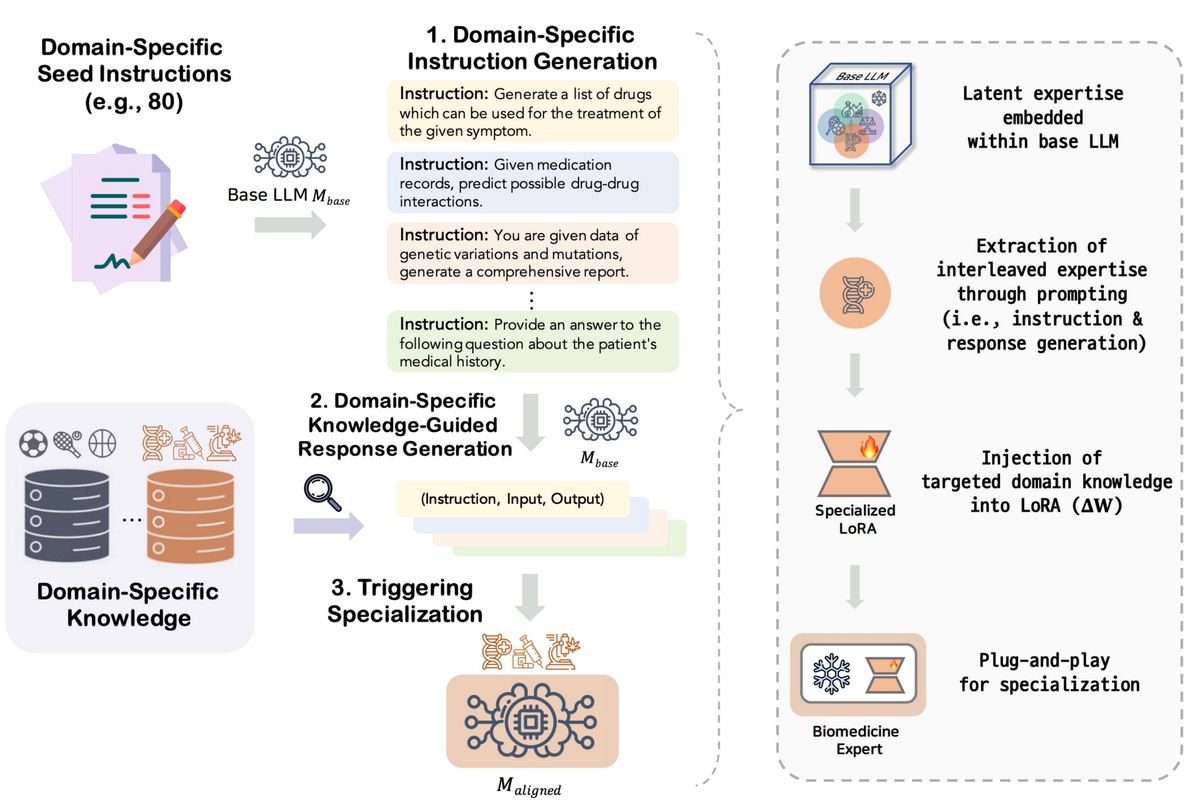
"Aligning to handle a broad set of instructions with only minimal supervision" - **Great!**
Do these generally aligned LLMs also generalize well in specialized domains ?





Only a slight advantage over the base model, although they are aligned to handle a broad set of instructions

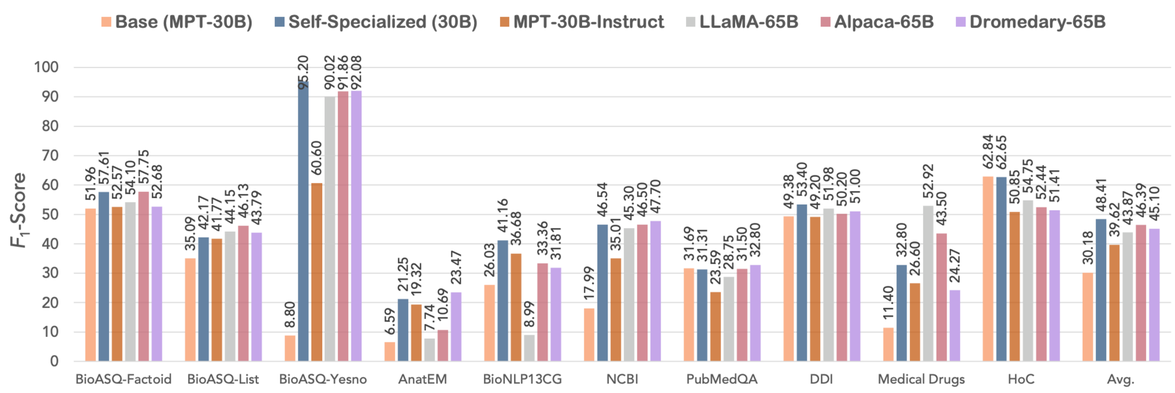
Self-Specialization for Uncovering Domain Expertise



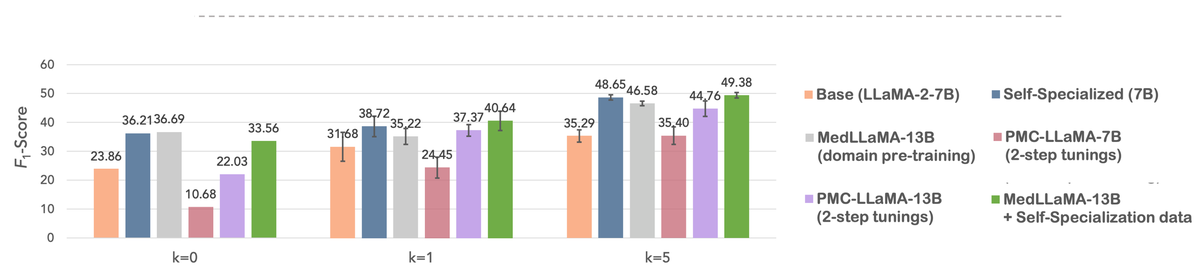
(a) Self-Specialization Process

(b) Abstraction of Process

Results in Biomedical Domain

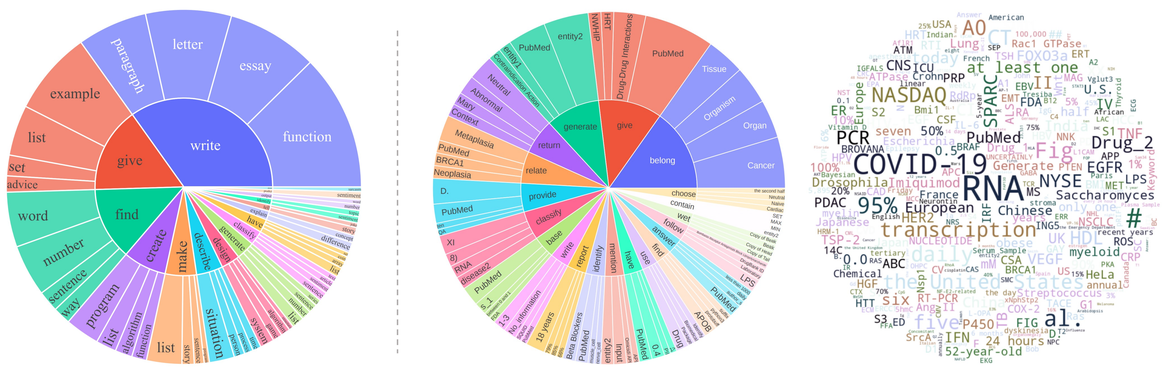


Self-specialization significantly improves its base model (up to 18 points!), even surpassing 65B models



Self-specialization is on par or more effective than extensive domain pre-training + complementary

Generated Data through Self-Specialization

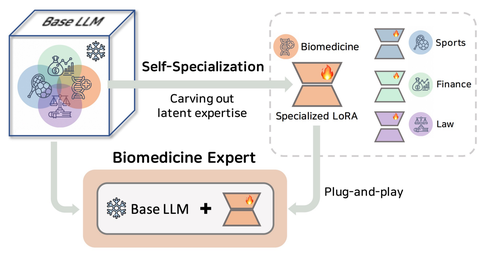


General (from Self-Instruct)

Specialized (Target: Biomedicine)

Key Takeaways

Q. Can we self-align LLMs with an expert domain like biomedicine with limited supervision?



1. **Benchmarking of General-Purpose Aligned Models**
   Highlighting the intrinsic challenge of encoding vast general knowledge into a finite set of parameters
2. **Exploring a Lightweight Solution, Self-Specialization**
   Targeted self-alignment to uncover latent expertise within LLMs with minimal supervision
3. **Findings**
   - Remarkable effectiveness in biomedical and financial domains
   - Highly efficient and practical: Tuning with QLoRA on single A100 (using 5K generated data, ~3 hrs)