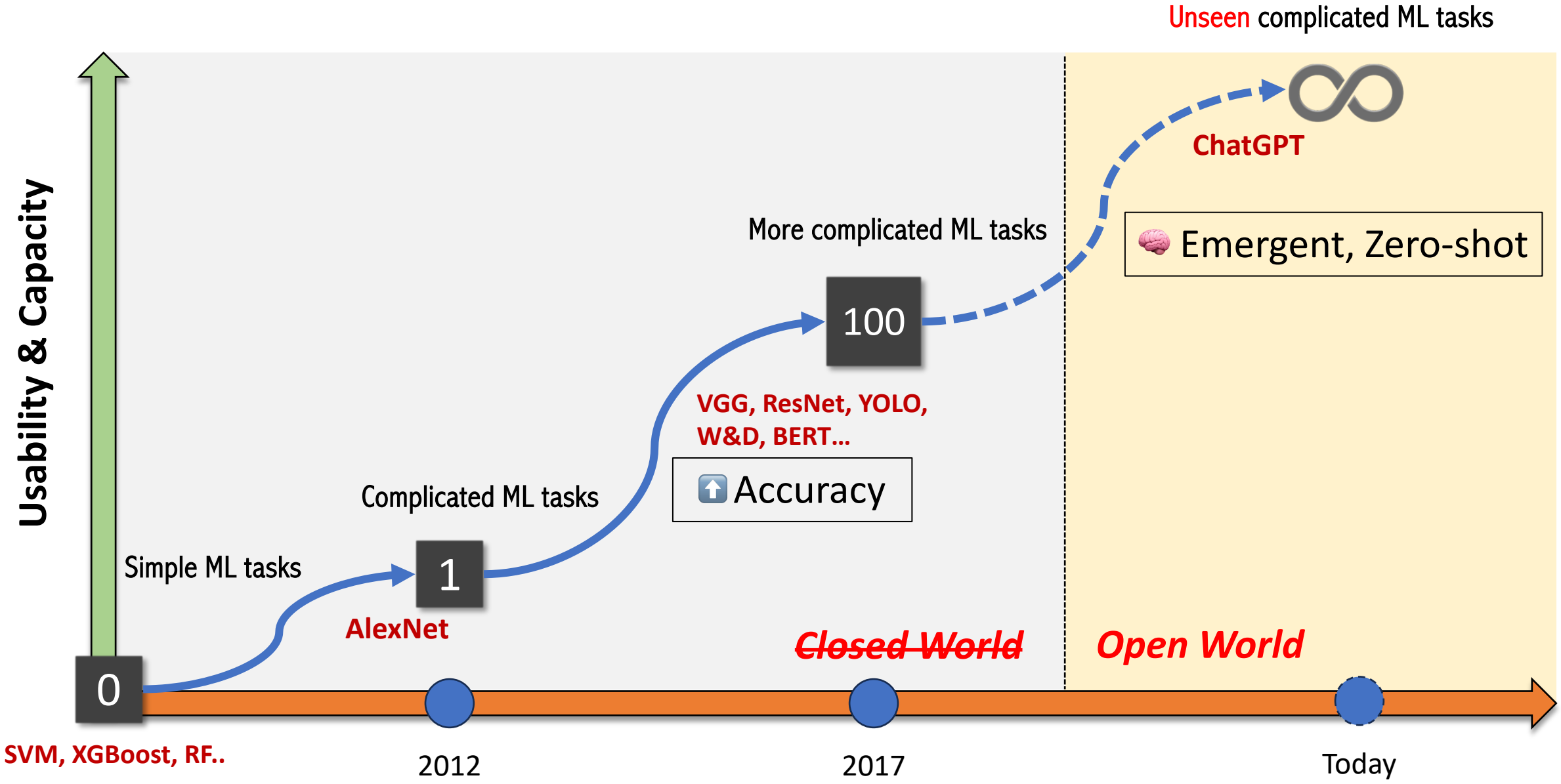
What ChatGPT means to AI..

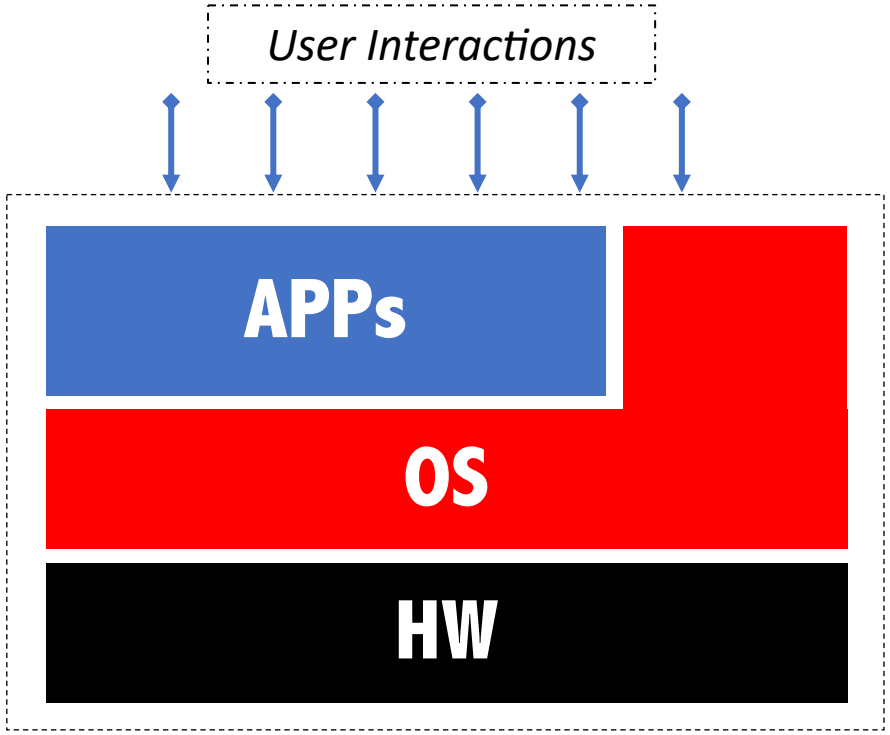
- “ChatGPT is just a smarter chatbot”
  - As a product, yes
    - But think about it: Moss is also a chatbot; robots/humans are chat bots with physical ability
  - As a research, hell no
    - It is a generative model that theoretically knows everything on Internet and can accomplish any NLP tasks
  - It’s also
    - a series of papers cited by 10,000 times
    - a startup company worthy of 30,000,000,000 dollars.
  - It’s also the one who opens the Pandora’s box

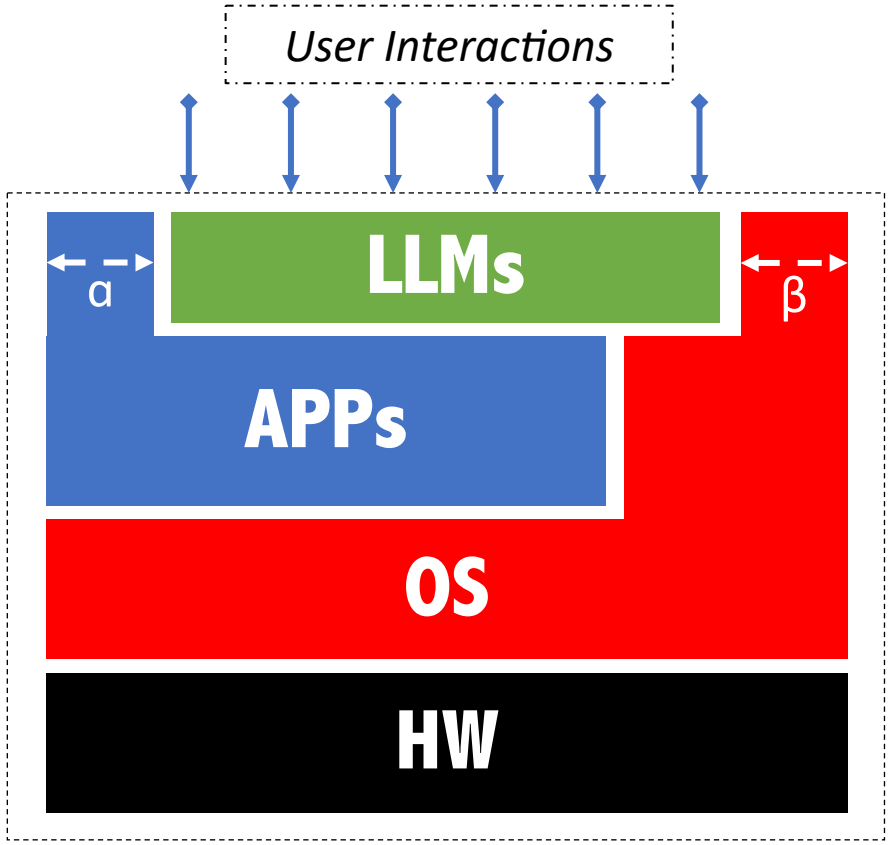
What LLM means to AI..



LLM is the new Operating System

- Users interact with LLM, while LLM manages/utilizes old-time apps/OS and hardware





α+β -> 0

LLM and Operating Systems

- LLM + OS: where to go?
  - How intrusive LLM shall be into the old and complex Oses
  - LLM is the new chance for new OSes (another Linux?)

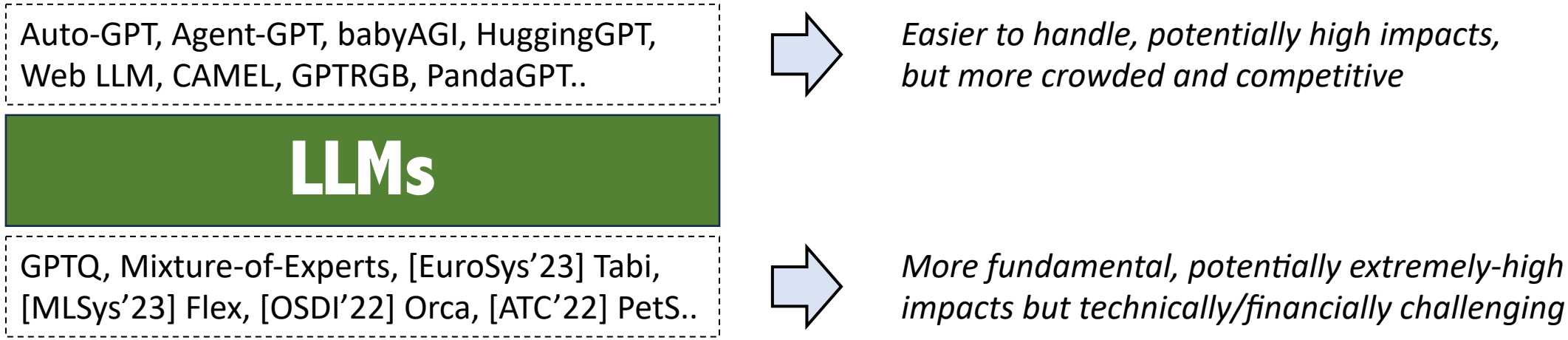
1. **Non-intrusive**: LLM behaves as an agent; it invokes system APIs and UI operations on existing OSes
2. **Half-intrusive**: re-engineer the OSes to better fit to LLM
   - Top-down approach, LLM as developer with JIT
   - Mostly the syscall interfaces and user-space libs
3. **Full-intrusive**: building new OSes (with LLMs) for LLMs
   - Bottom-up approach, LLM as the god
   - Modularity and expressiveness are the keys

The New Golden Era for Mobile Research

- Since iPhone 2007..
- The next long-term goal of mobile research: ChatGPT on smartphone
  - Takes 5 ~ 10 years
  - Takes collective efforts from hardware/architecture, mobile system, ML algorithm communities
  - LLM on edge vs. LLM-as-a-Cloud-Service
- **Old stories:** data privacy, low delay, low power consumption, etc..
- **New techniques:** memory-bounded LLMs, foundation model + adapters, generative and augoregressive, etc..

Exploration Atop or Below LLM?

- Another way to go: build systems for LLM, or build systems with LLM
- When a software layer is finalized, most research/industry opportunities go above
  - Very very few system researchers rebuild OS now
  - Very very few network researchers rebuild network stacks now



A Paper list for resource efficient LLM: https://github.com/UbiquitousLearning/Paper-list-resource-efficient-large-language-model