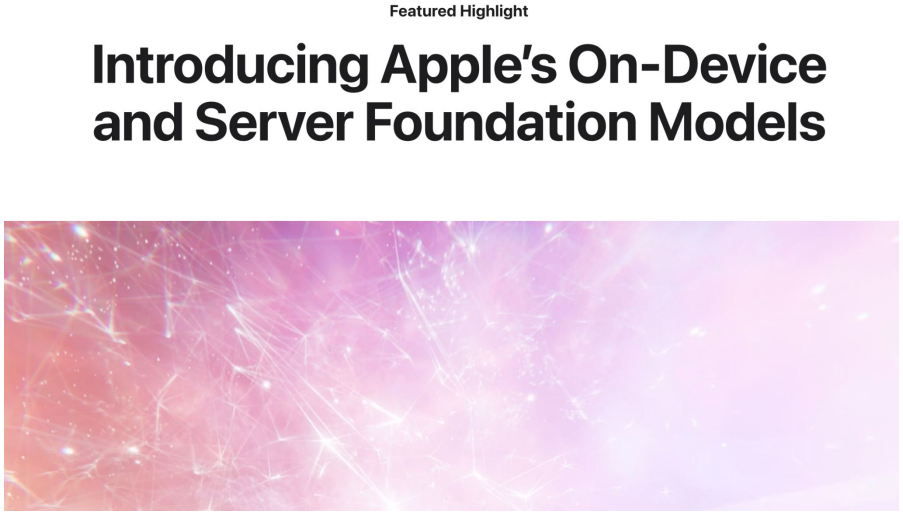
Apple’s New On-device and Server Foundation Models

Roger Lam
lamroger.com

Apple published an overview on two of their new LLMs

- 3B parameter model for iPhones named **Apple On-Device**
- Undisclosed parameter larger model named **Apple Server**
- They are fine-tuned for user experiences (writing, summarizing, playful images, actions)
- https://machinelearning.apple.com/research/introducing-apple-foundation-models

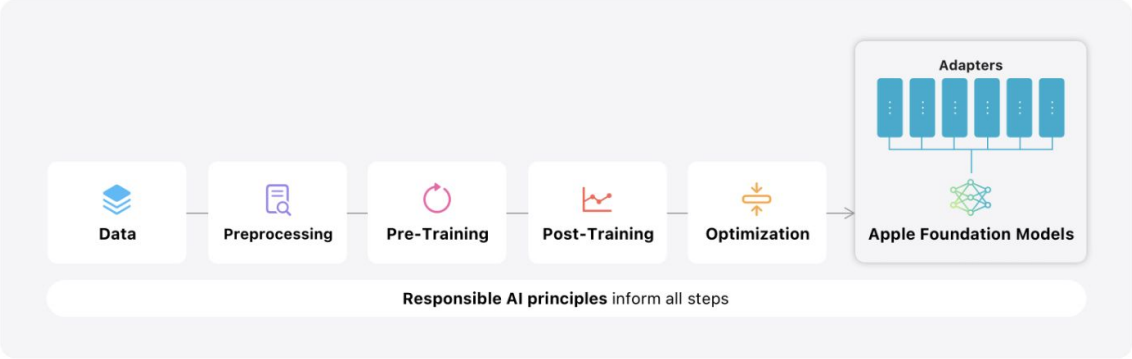


Apple details their Responsible AI principles

- You can read them on the right but they focus on the user first, acknowledge that the technology take a lot of continual testing, and privacy is non-negotiable
- Nothing surprising imo

1. **Empower users with intelligent tools**: We identify areas where AI can be used responsibly to create tools for addressing specific user needs. We respect how our users choose to use these tools to accomplish their goals.
2. **Represent our users**: We build deeply personal products with the goal of representing users around the globe authentically. We work continuously to avoid perpetuating stereotypes and systemic biases across our AI tools and models.
3. **Design with care**: We take precautions at every stage of our process, including design, model training, feature development, and quality evaluation to identify how our AI tools may be misused or lead to potential harm. We will continuously and proactively improve our AI tools with the help of user feedback.
4. **Protect privacy**: We protect our users' privacy with powerful on-device processing and groundbreaking infrastructure like Private Cloud Compute. We do not use our users' private personal data or user interactions when training our foundation models.

Training uses TPUs and GPUs, licensed and crawled data



- I didn’t expect Apple to use TPUs
- They also use crawled data with AppleBot - transparency is good but it can still crawl copyrighted data

Small but maybe meaningful strategies in Post-Training

- They use high quality human-annotated and synthetic data
- Reminds me of Phi models using “textbooks”

Small but maybe meaningful strategies in Post-Training

- They also detail “two novel algorithms in post-training”
- A **rejection** sampling fine-tuning algorithm **with teacher committee**
  - Multiple specialized models to filter data
- A reinforcement learning from human feedback (RLHF) algorithm with **mirror descent policy optimization** and a **leave-one-out advantage estimator**
  - MDPO - “mirror” what the existing value is by not changing too much
  - LOO - see what the overall value is if you leave out the current one
- My first time hearing of both strategies. LOO comes from traditional ML techniques.
- It seems like we’re beginning to apply known techniques to overcome challenges

Their optimization team is working hard

- Two measures are time-to-first token and tokens per second
- Grouped query attention,
- shared embedding tables in input and output
  - Since tables are shared, we only need to map once with encoder / decoder
  - Is it an encoder / decoder model?
- Quantize intelligently - “low-bit palletization”
  - They still use LoRA adapters but mix 2-bit and 4-bit configs, **avging 3.5 bits-per-weight**, **“achieving the same accuracy as the uncompressed model”**
  - Parameter palletization = grouping and sharing similar parameters
  - And activation quantization, embedding quantization, efficient KV updates

Their optimization team is working hard

- Result on iPhone 15 Pro: **time-to first-token is 0.6ms per prompt token** and **generation rate of 30 token per second**.
  - Using GPT-4 tokenizer estimator and the contents of the blog post, it's about 5,200 tokens or about 3 seconds til first token
- 3 seconds isn't bad for analyzing a website. I imagine they'll continue to work to drop this down incrementally.
- Token Estimator - https://platform.openai.com/tokenizer
- Token per second simulator - https://tokens-per-second-visualizer.tiiny.site/

Model Adaptation aka LoRA Adaptors

- They develop small adaptors for individual tasks or properties
- Each are 10s of MBs
- **Do they stack? Probably not?**

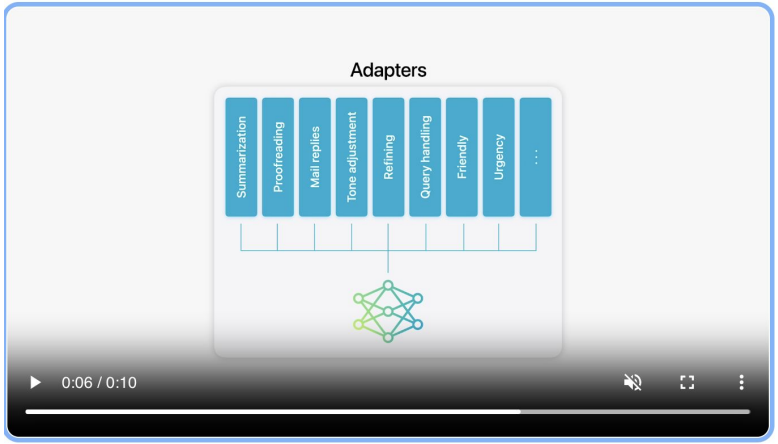
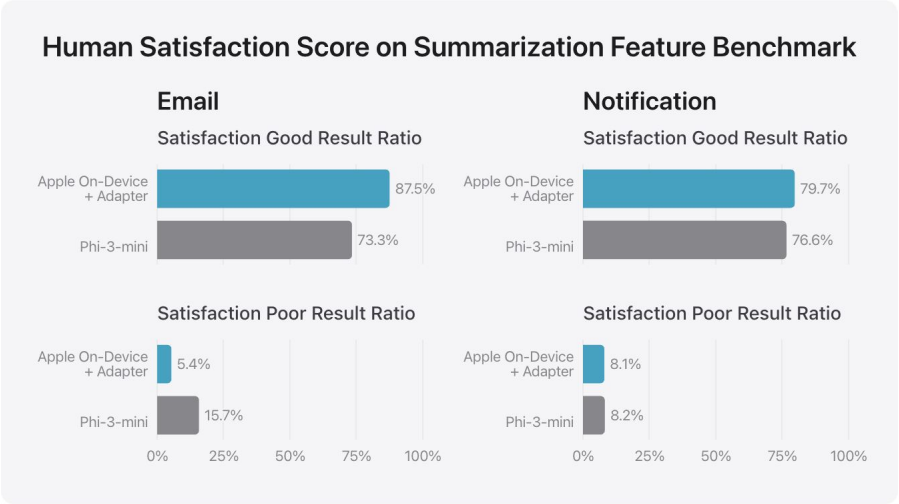


Figure 2: Adapters are small collections of model weights that are overlaid onto the common base foundation model. They can be dynamically loaded and swapped — giving the foundation model the ability to specialize itself on-the-fly for the task at hand. Apple Intelligence includes a broad set of adapters, each fine-tuned for a specific feature. It's an efficient way to scale the capabilities of our foundation model.

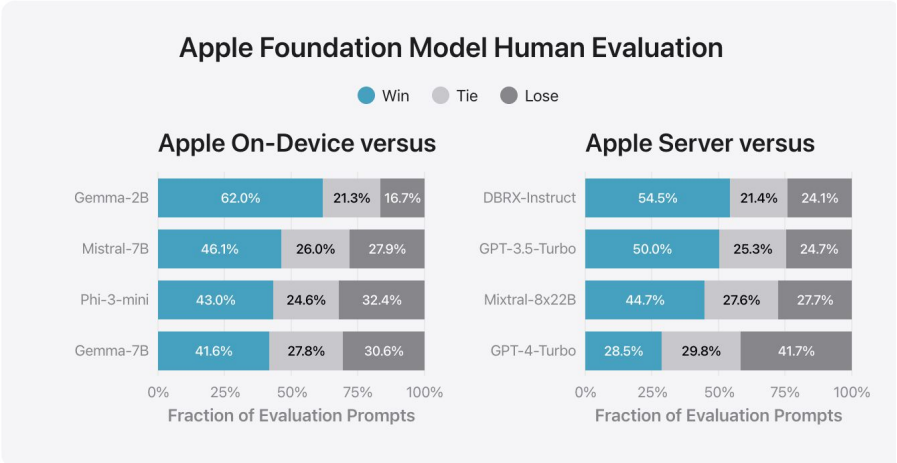
Benchmarks are slightly better than Phi-3-mini (3.8B)

- They use their own benchmarks, a set of 750 responses for each use case
- **WITH adaptors**, Apple is better at email and notification.



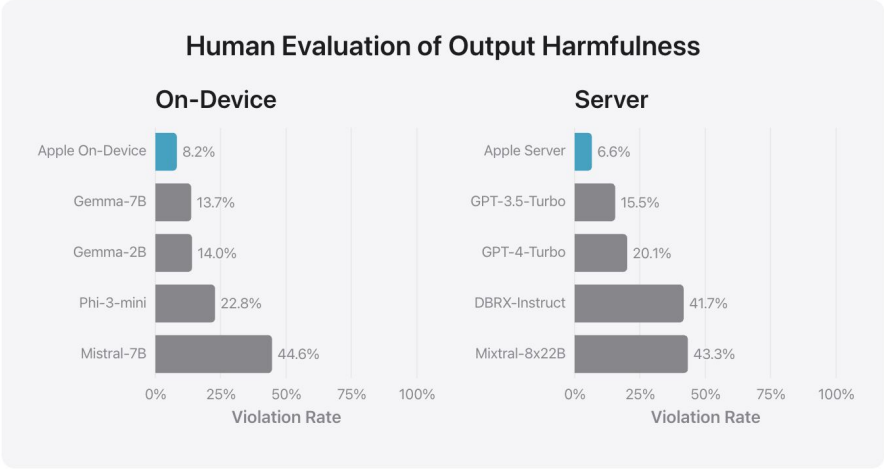
Human eval is comparable to similar models

- On-device is better vs other small models
- Apple server is better than GPT-3.5-turbo but not as good at GPT-4-turbo
  - No benchmarks against Llama, GPT-4o, Claude

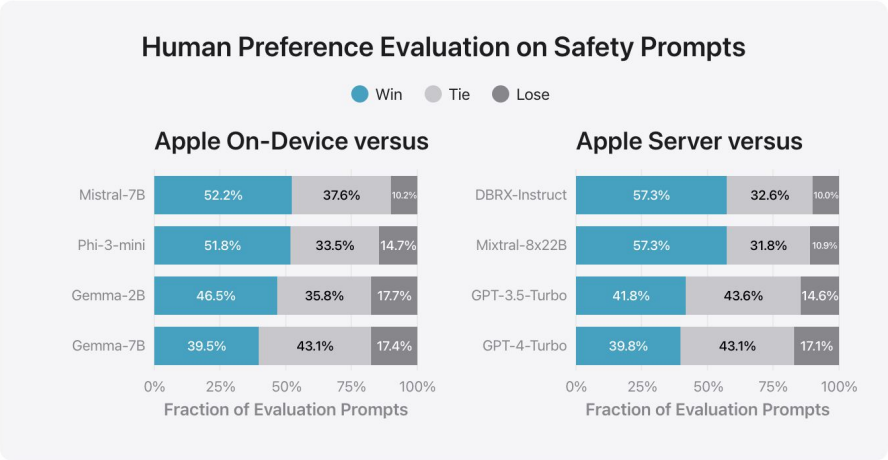


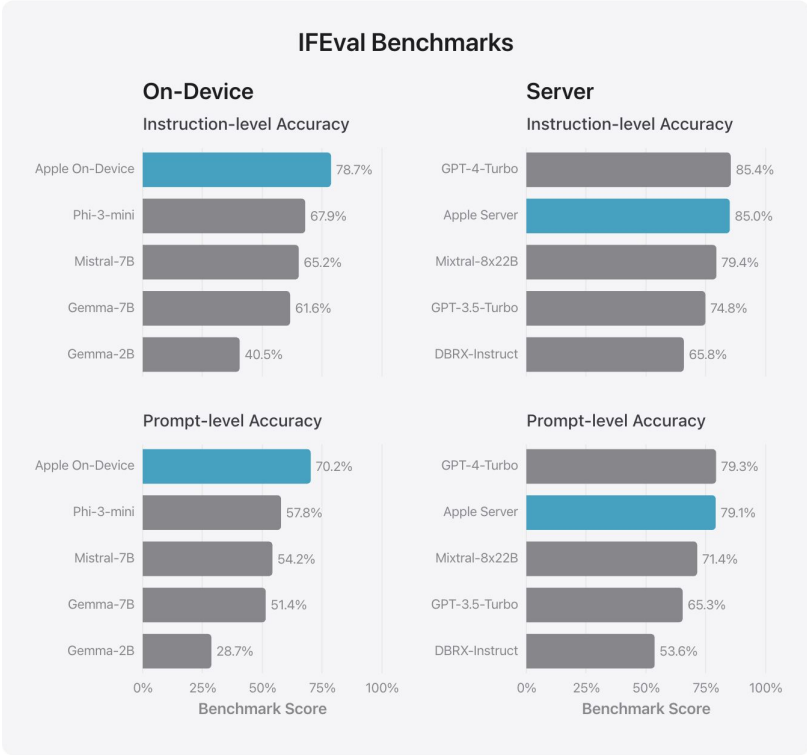
Harmfulness good but none again GPT-4o and Claude

- More guardrails and safety built in
- Notable omission against Anthropic, one that prides in being helpful, honest, and harmless



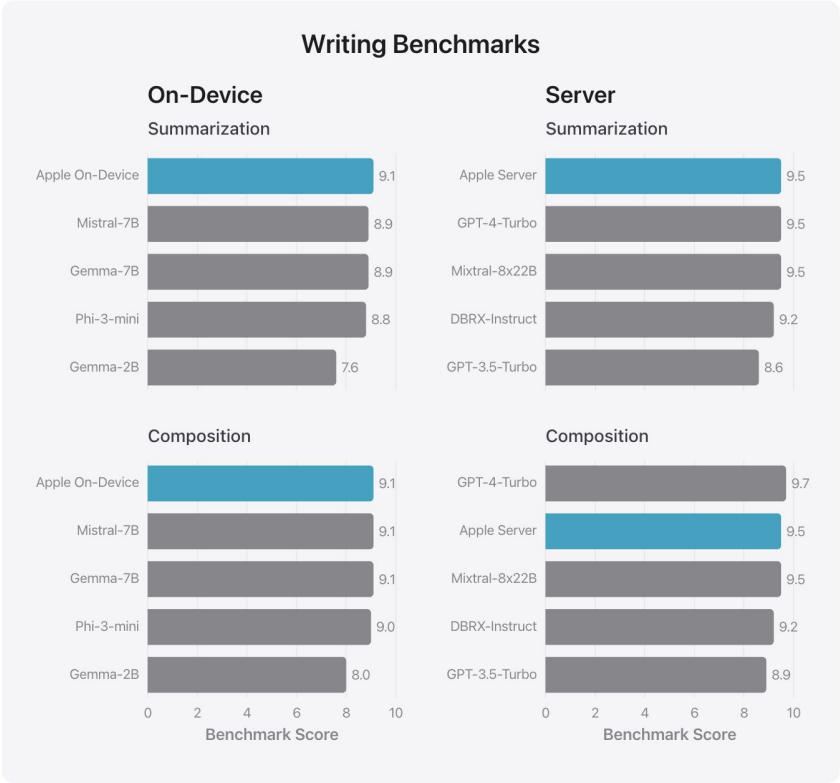
More on safety, instruction following, and writing





More on safety, instruction following, and writing

- They compared against open models and GPT-4-turbo which makes me put them at the same level.
- Could be that they didn’t want to leak comparisons against competitors while benchmarking
- I’m sure people will start benchmarking themselves when it become available.



Thanks!

- I’m going to keep these slides focused on just this blog post
- More soon!

www.lamorger.com

https://www.linkedin.com/in/lam-roger/