Head in the Cloud?

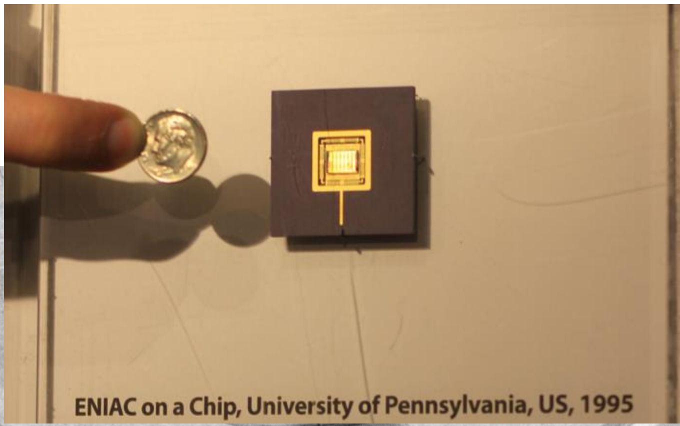
What’s on the Horizon for Offline Intelligence

Agenda

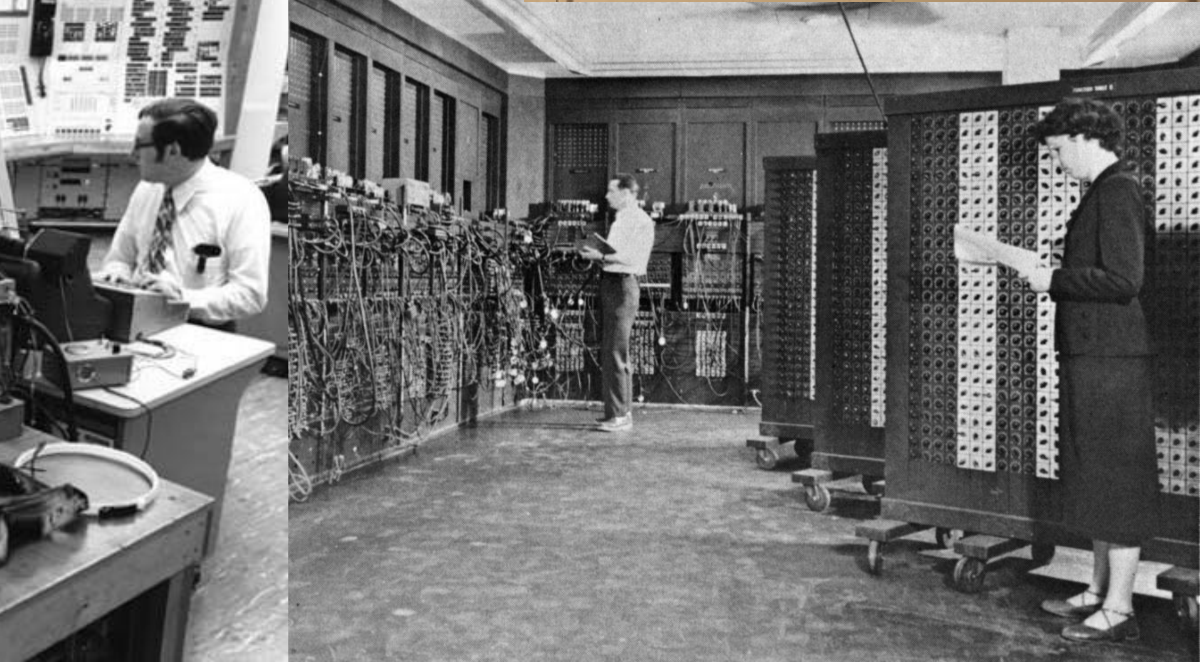
- Assistive Technology over time
  - Case Studies
- Characterizing Artificial Intelligence today
  - Case Studies
- Discussion

Flashback: Early Computers

From punch cards to terminals



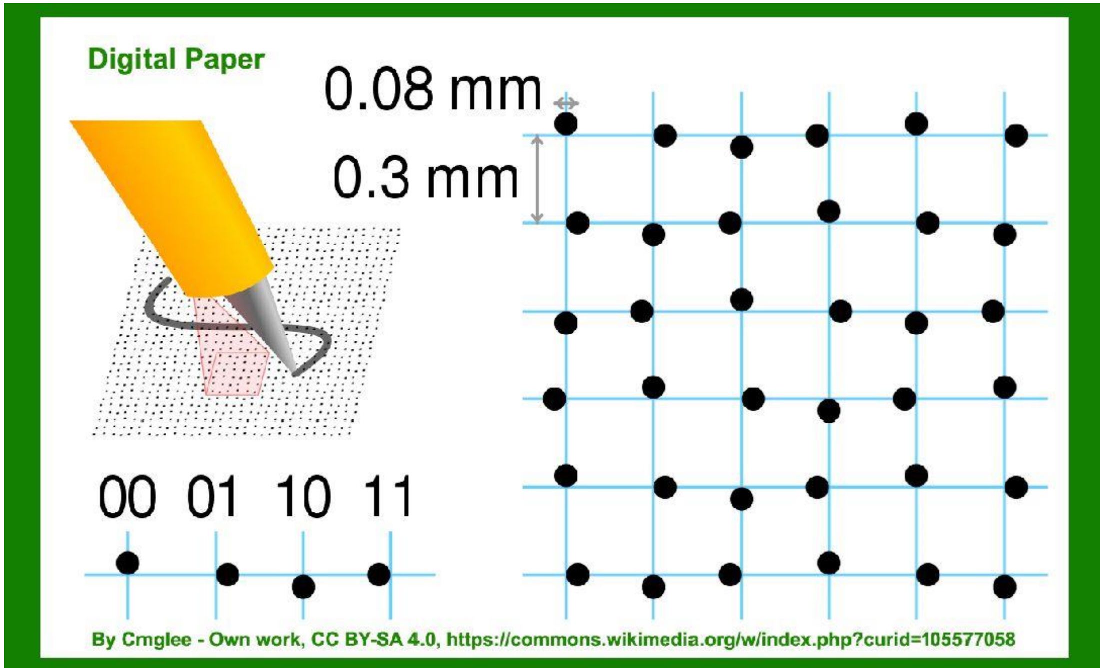
ENIAC on a Chip, University of Pennsylvania, US, 1995



Case Study: LiveScribe Smartpen

Complement handwritten notes with on-device audio using specialized pen and note paper

- Note paper & ink as the subscription component
- Built-in audio recording feature eventually offloaded to accompanying smartphone



Case Study: Voice Dream Reader

Popular Document Text-to-Speech tool (Mac, iOS)

Originally: one-time purchase at $29.99

Shifted to: monthly subscription of $9.99 or $49.99 annually

voice dream

Reader

Blog

Pricing Update for One-time Purchasers

[Update: April 6th]

Following our recent announcement to transition Voice Dream to a subscription, we received an overwhelming response from thousands in our community. Your feedback, along with the impactful stories shared about Voice Dream being a pivotal part of your daily lives, has led us to reverse this change.

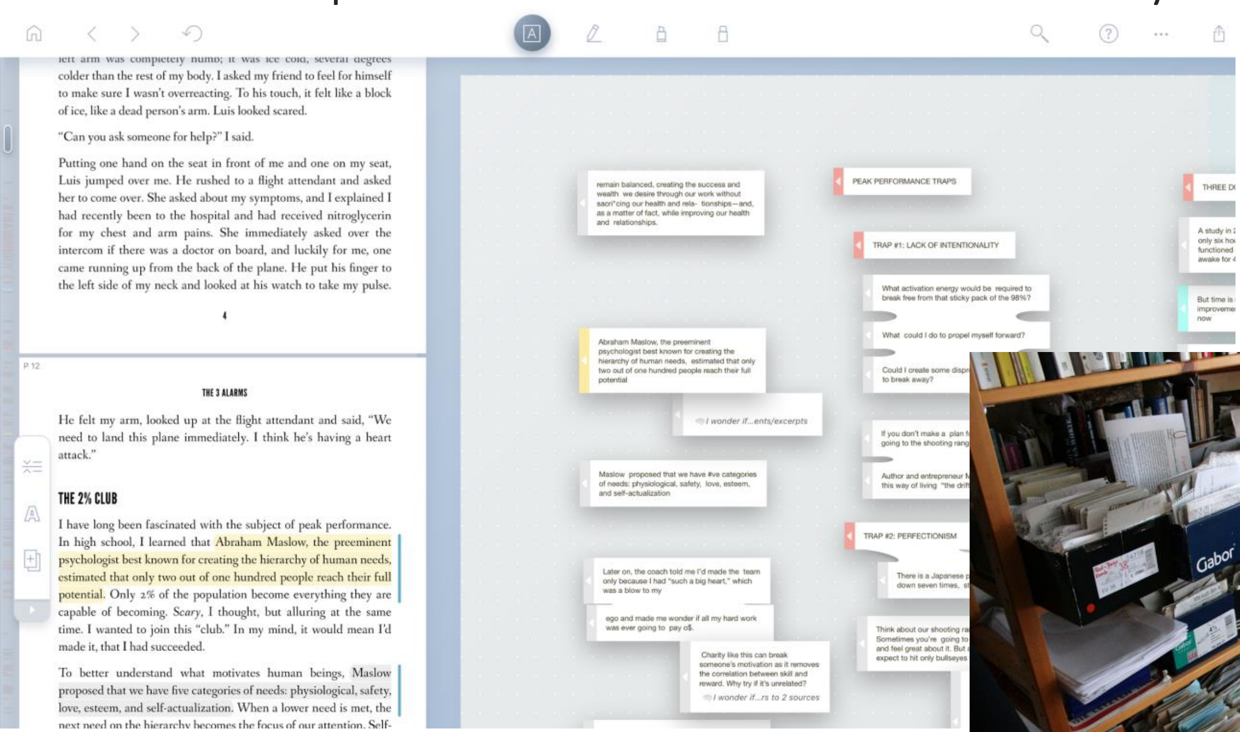
We will continue to provide access to the app's existing features at no additional cost.

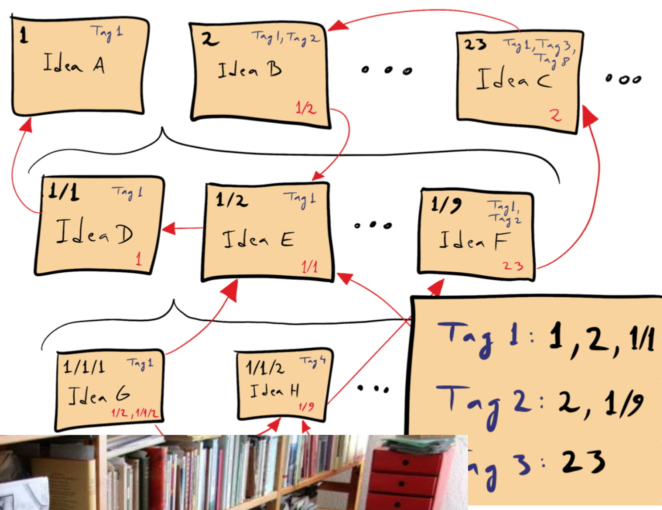
As we continue developing Voice Dream, some new features may be offered as part of a subscription, but the current capabilities will remain free to those who have already purchased Voice Dream.

Case Study: LiquidText

"Zettelkasten" PDF Note-taking tool (Windows, Mac, iPad)

- Offers version-specific, one-time purchase pricing
- Subscription unlocks cloud features & sync





Cloud-based App Experience

Yay!

- Bring your own device
- Instant access to latest technology
- Continuous development
- Competition
- Low entry cost

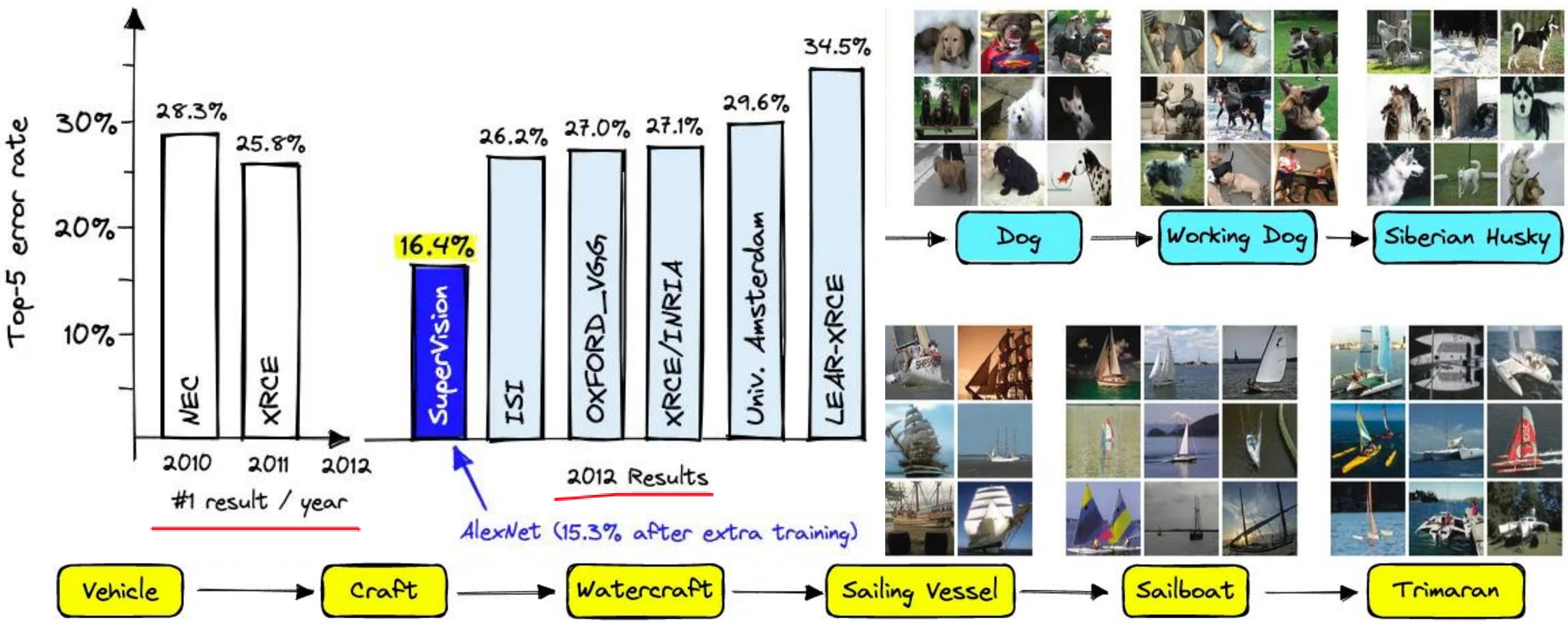
Ugh!

- One interface per app
- Feature bloat
- Recurring, often increasing per-seat fees
- Introductory vs renewal pricing
- Complicated privacy and accessibility evaluations subject to changes over time

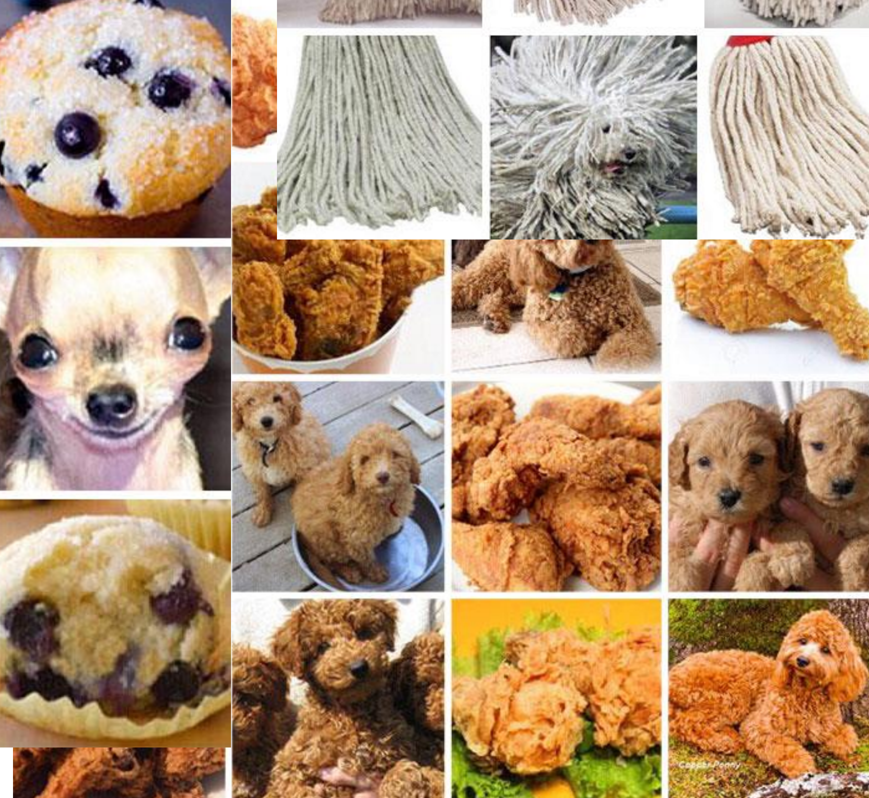

Flashback: Kurzweil Reading Machine

Artificial Intelligence (AI) Primer

ImageNet Database and associated competition

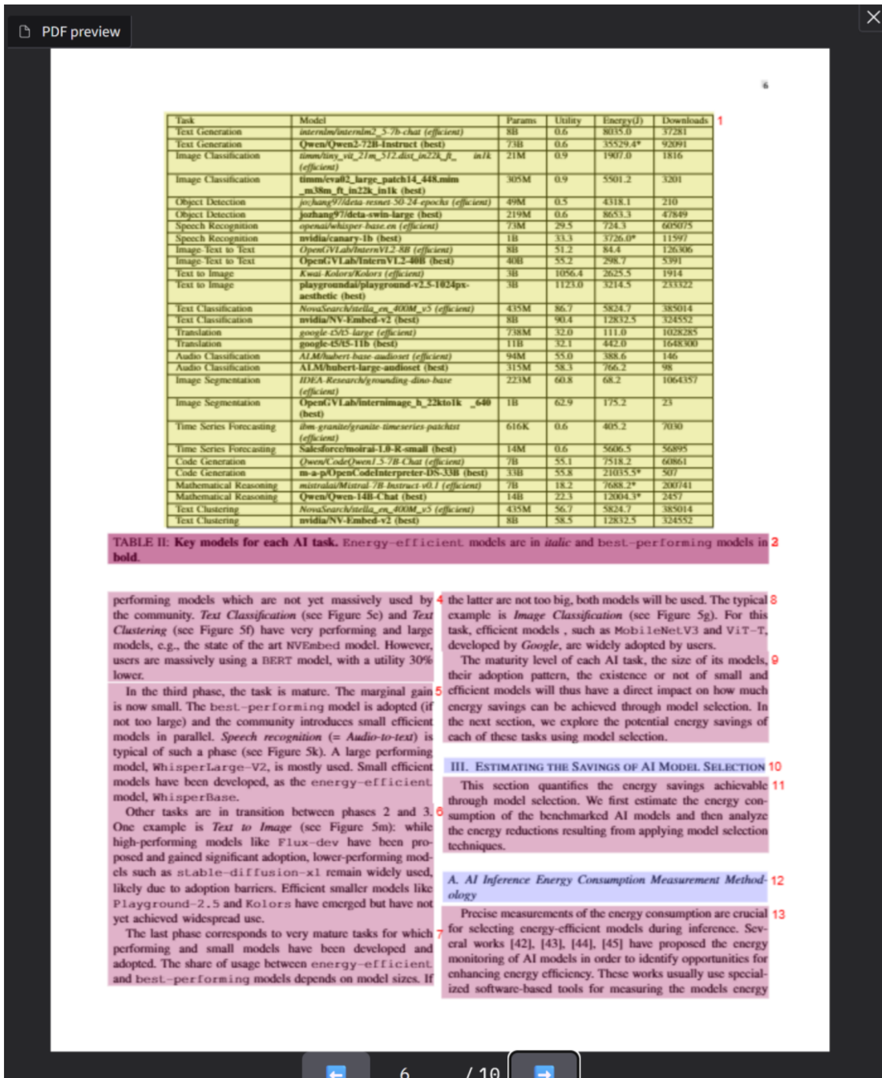


Let’s check it out!



Offline slide

Case Study: MinerU



Text Clustering	nvidia/NV-Embed-v2 (best)	8B	58.5	12832.5	324552

TABLE II: Key models for each AI task. Energy-efficient models are in italic and best-performing models in bold.

TABLE II: Key models for each AI task. Energy-efficient models are in italic and best-performing models in bold.

performing models which are not yet massively used by the community. Text Classification (see Figure 5e) and Text Clustering (see Figure 5f) have very performing and large models, e.g., the state of the art NVEmbed model. However, users are massively using a BERT model, with a utility 30% lower.

In the third phase, the task is mature. The marginal gain is now small. The best-performing model is adopted (if not too large) and the community introduces small efficient models in parallel. Speech recognition (= Audio-to-text) is typical o such a phase (see Figure 5k). A large performing model, WhisperLarge-V2, is mostly used. Small efficient models have been developed, as the energy-efficient model, WhisperBase.

Other tasks are in transition between phases 2 and 3. One example is Text to Image (see Figure 5m): while high-performing models like Flux-dev have been proposed and gained significant adoption, lower-performing models such as stable-diffusion-x1 remain widely used, likely due to adoption barriers. Efficient smaller models like Playground-2.5 and Colors have emerged but have not yet achieved widespread use.

The last phase corresponds to very mature tasks for which performing and small models have been developed and adopted. The share of usage between energy-efficient and best-performing models depends on model sizes. If

the latter are not too big, both models will be used. The typical example is Image Classification (see Figure 5g). For this task, efficient models, such as MobileNetV3 and ViT-T, developed by Google, are widely adopted by users.

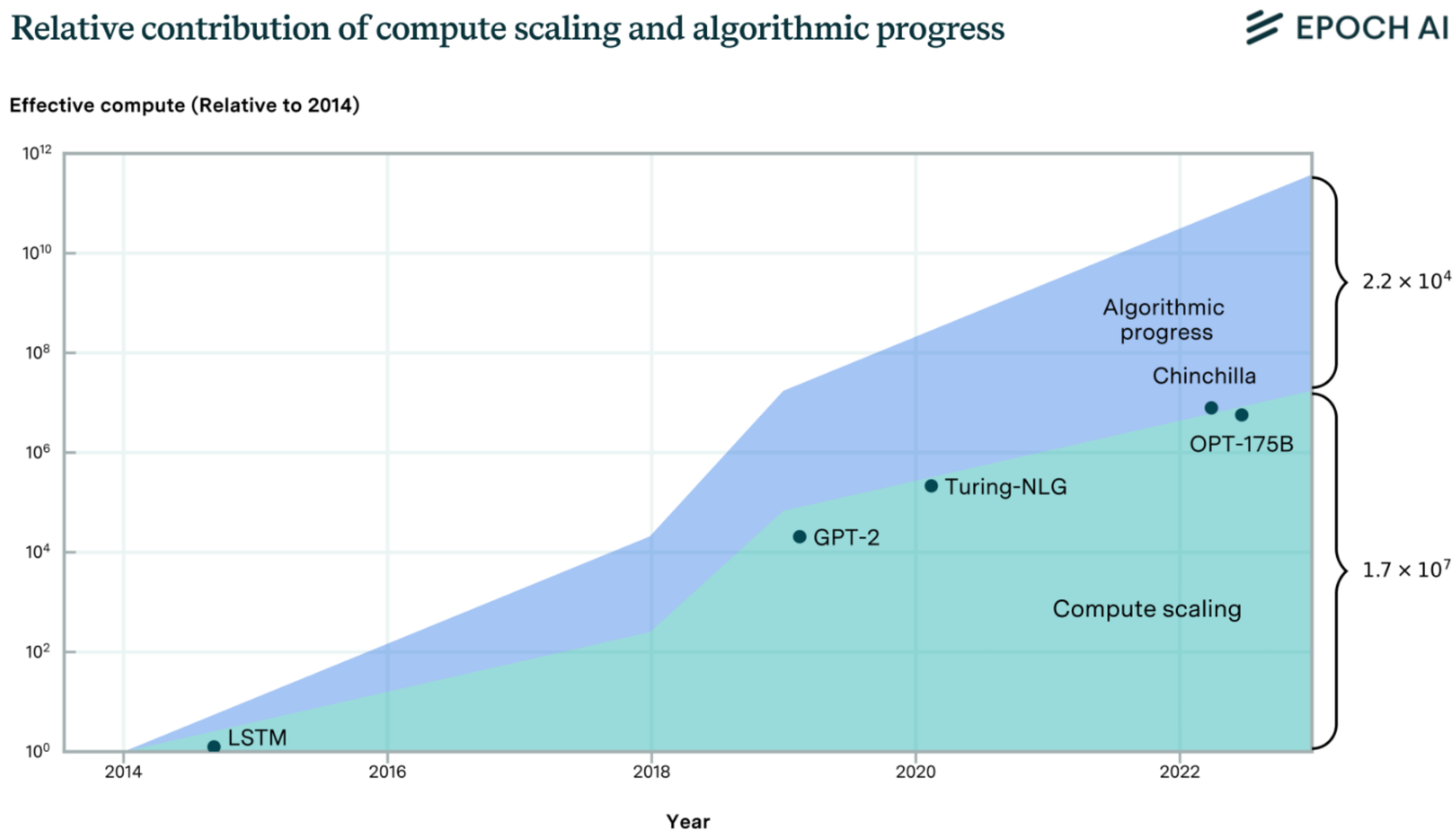
The maturity level of each AI task, the size of its models, their adoption pattern, the existence or not of small and efficient models will thus have a direct impact on how much energy savings can be achieved through model selection. In the next section, we explore the potential energy savings of each of these tasks using model selection.

III. ESTIMATING THE SAVINGS OF AI MODEL SELECTION

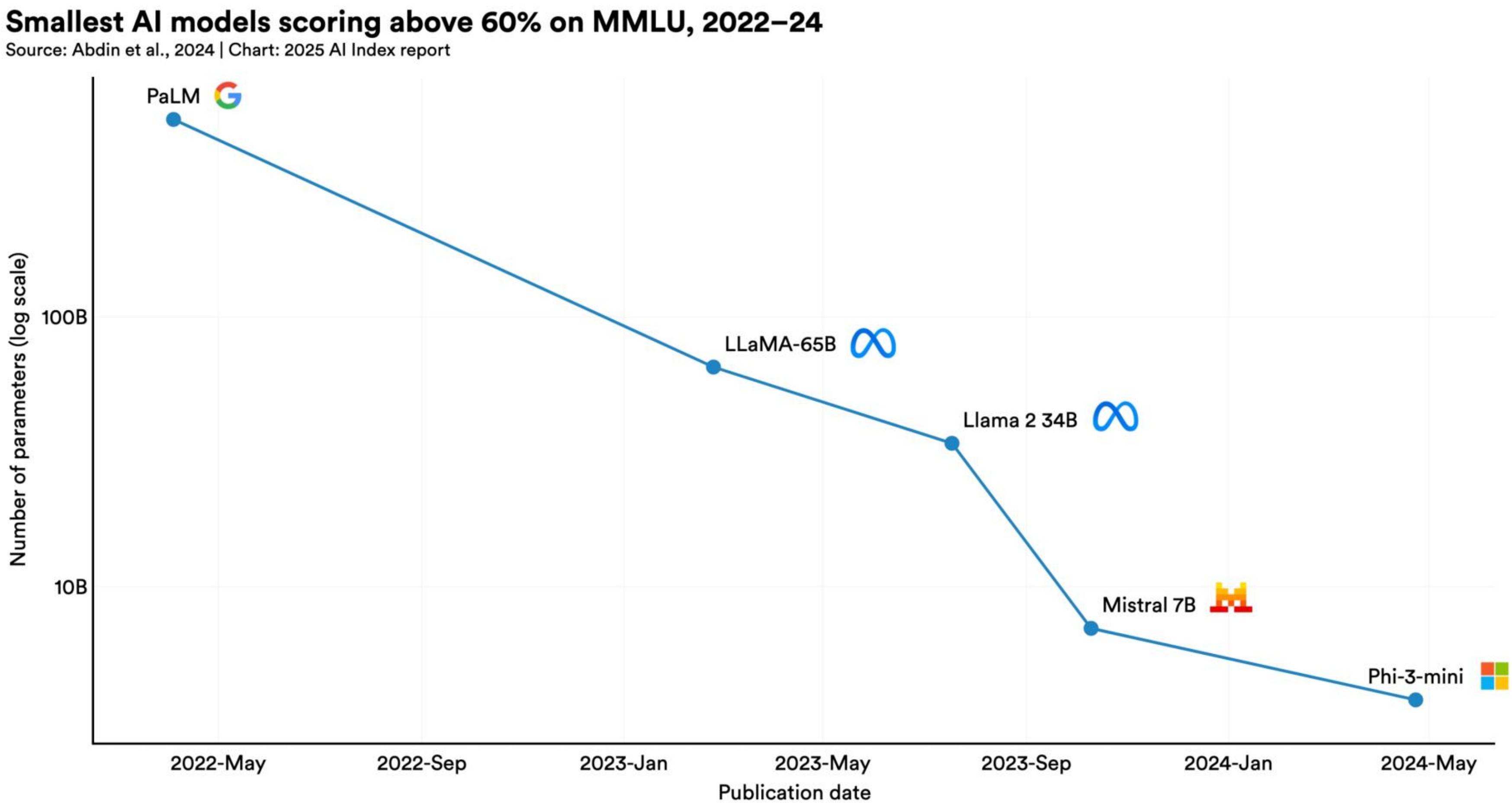
This section quantifies the energy savings achievable through model selection. We first estimate the energy consumption of the benchmarked AI models and then analyze the energy reductions resulting from applying model selection techniques.

Transformer Architecture

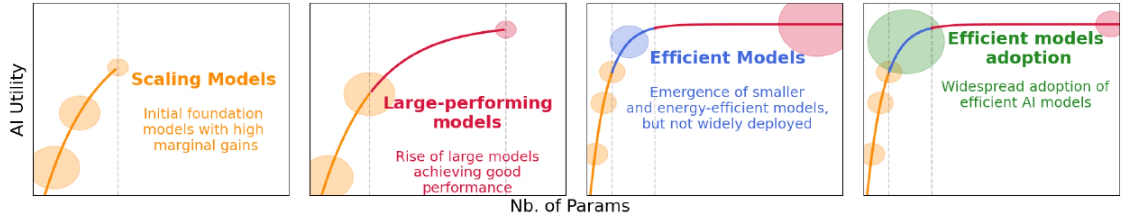
Full context -- Massive scale-up of parallel computation



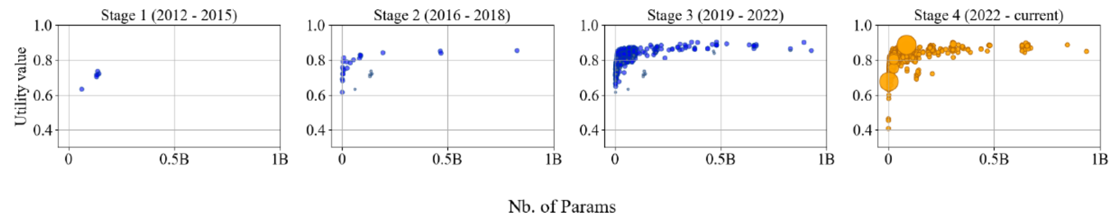
Not just Bigger and Better

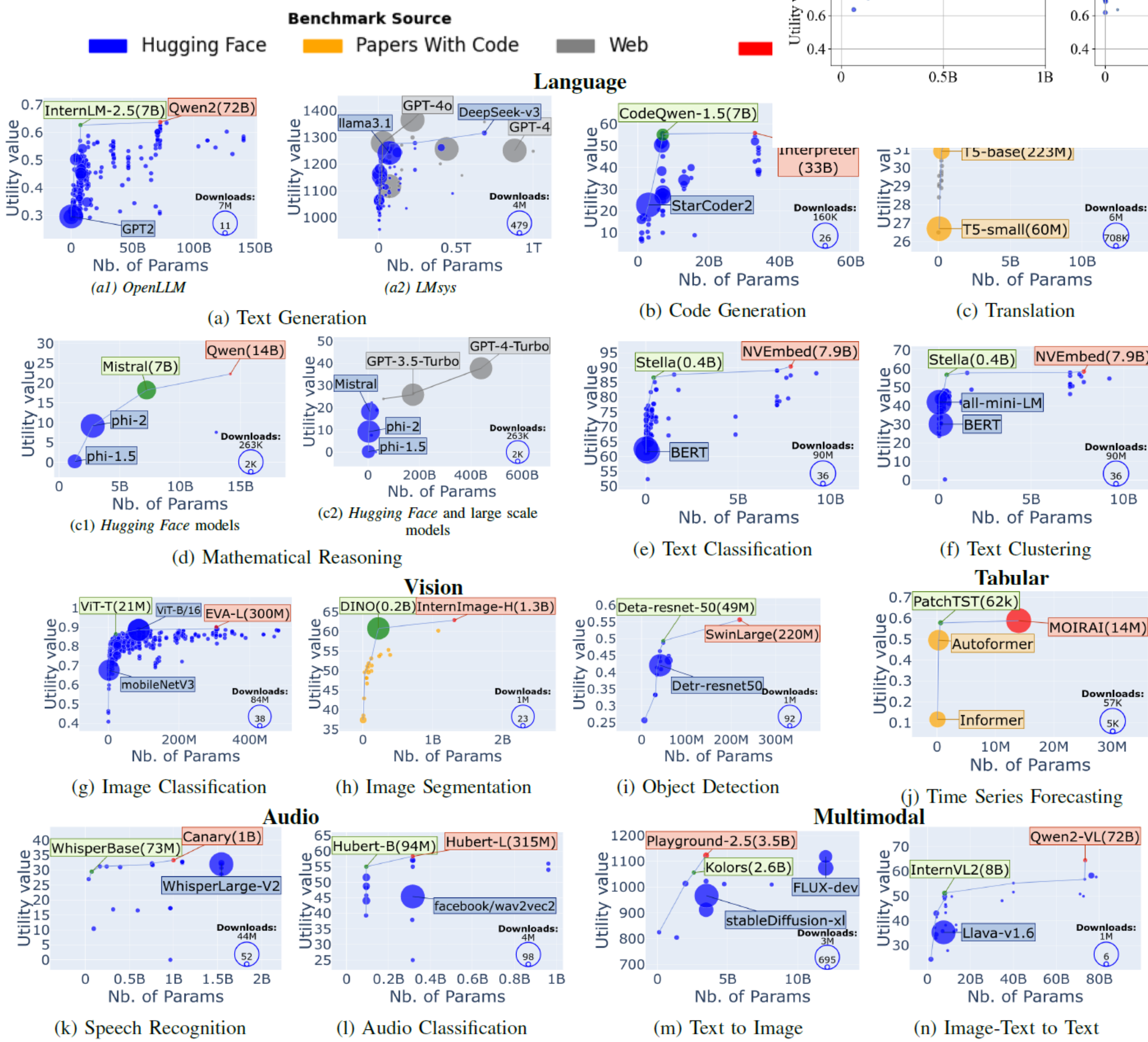


Small is Sufficient



(a) The 4 stages of development of an AI task over time





Case Study: Apple Spoken Content

Text-to-speech feature available on Apple devices

- Long-standing accessibility feature for offline playback of on-screen text with downloadable voices
- 30+ languages but premium voices limited to small handful
- Once downloaded, a voice becomes available for other text to speech apps
- iPhone 12/Apple Silicon: Apple Notes transcription available offline for select languages
- *Shoutout to Live Speech including offline Personal Voice*

Case Study: Google Pixel Recorder App

Speech to text app specific to Google Pixel phones

Works entirely offline using specialized hardware and AI-model

First became available packaged with the Pixel 5 (released in 2020)

Includes advanced features such as speaker identification

English-language only

Shoutout to Google AI Edge Gallery for exploring on-device AI capabilities

Case Study: Windows Recall

Chat with snapshots of your screen going back in time (Windows Copilot+ PCs only)

- Offline after one-time download of AI components
- Concerns about privacy incl. inconsistent filtering of sensitive information
- Dependent on user education around local security
- Limited to 6 primary languages
- *Shoutout to Windows 11 Live Captions (plus translation from/into dozens of languages on Copilot+ PCs)*

Discussion

Takeaways

The future of collaborative learning requires real-time responsiveness favoring offline AI

Consider the implications of training students on hardware they already own as opposed to on software they will perhaps only access during their time at your college