

# Head in the Cloud?

What's on the Horizon for Offline Intelligence

# Agenda

Assistive Technology over time

Case Studies

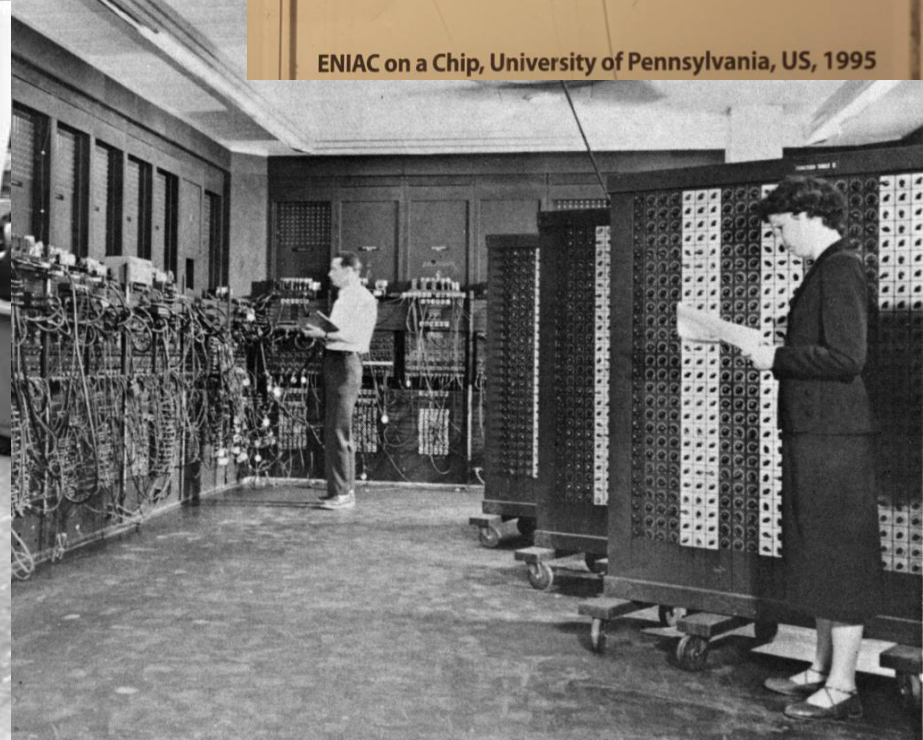
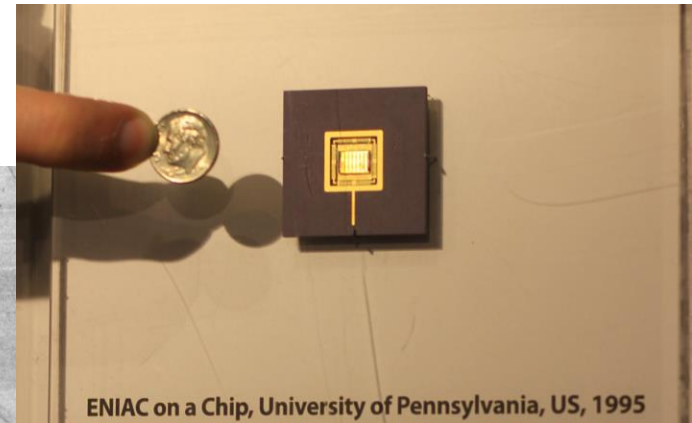
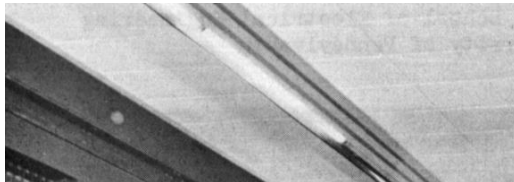
Characterizing Artificial Intelligence today

Case Studies

Discussion

# Flashback: Early Computers

From punch cards to terminals

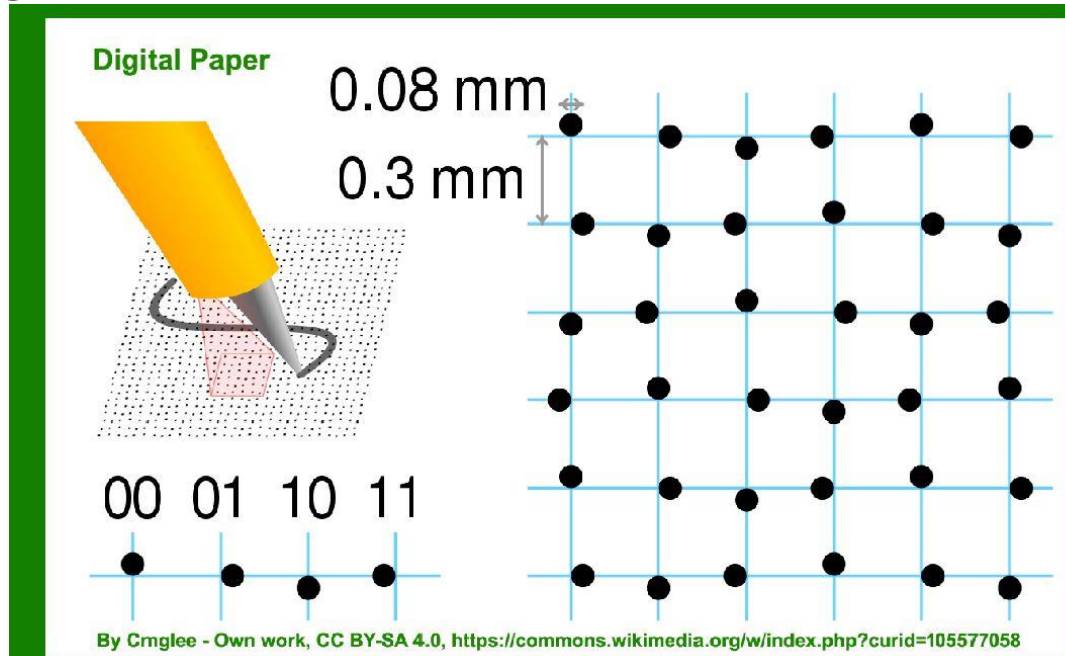


# Case Study: LiveScribe Smartpen

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.

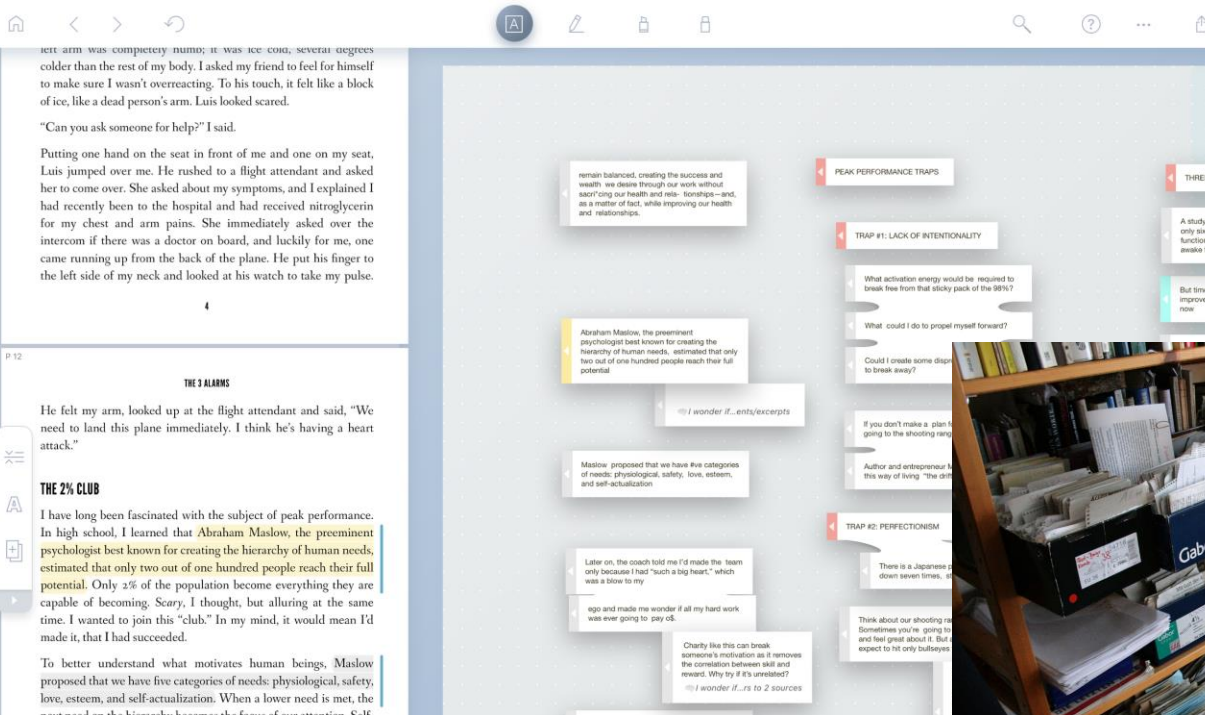
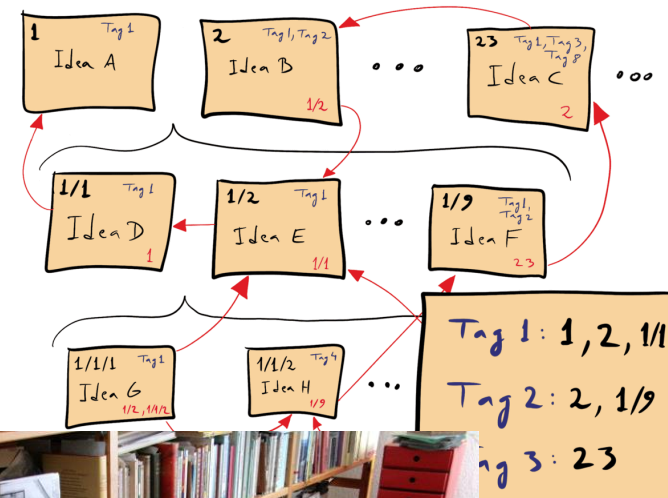
As we continue developing Voice Dream, some new features may be offered as part of a subscription, but the current capabilities will remain free to those who have already purchased Voice Dream.

# 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



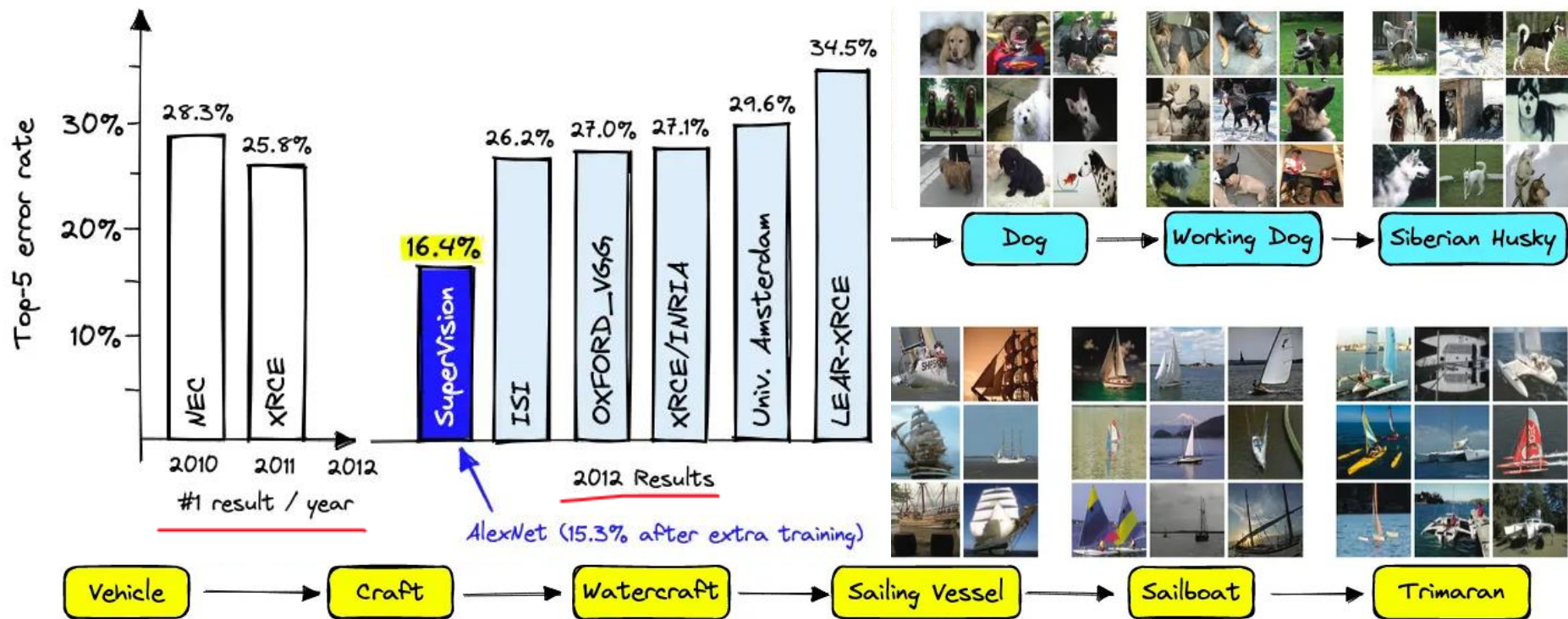


# Flashback: Kurzweil Reading Machine



# Artificial Intelligence (AI) Primer

## ImageNet Database and associated competition

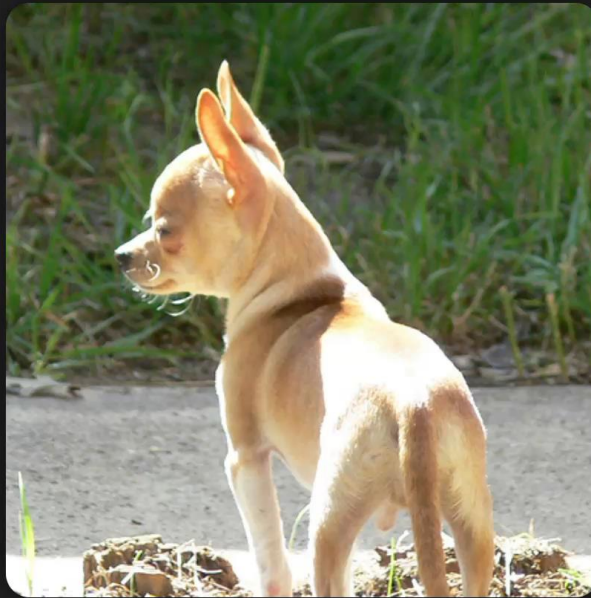




# Let's check it out!



# Offline slide



AI Guess: Chihuahua (45.2%) | dingo, warrigal, warragal, Canis dingo (21.5%) | basenji (6.4%)



# Case Study: MinerU

PDF preview

Task	Model	Params	Utility	Energy(J)	Downloads
Text Generation	<i>anyscale/internlm2.5-7b-chat (efficient)</i>	8B	0.6	8035.0	37281
Text Generation	<b>Qwen/Qwen2.72B-Instruct (best)</b>	73B	0.6	35529.4*	92591
Image Classification	<i>timonwly/vit_21m_312k_dist_patch16_in1k (efficient)</i>	21M	0.9	1907.0	1816
Image Classification	<b>timonwly/21m_312k_dist_patch16_in1k (best)</b>	305M	0.9	5501.2	3201
Object Detection	<i>ju-hang97/dota-reinet-50-24-epochs (efficient)</i>	49M	0.5	4318.1	210
Object Detection	<b>ju-hang97/dota-reinet-50-24-epochs (best)</b>	219M	0.6	8653.1	47849
Speech Recognition	<i>openai/whisper-base-en (efficient)</i>	73M	29.5	724.3	659375
Speech Recognition	<b>openai/whisper-base-en (best)</b>	1B	33.3	3726.0*	11597
Image-text to Text	<i>OpenGVLab/InternV1.2-8B (efficient)</i>	8B	51.2	84.4	126306
Image-text to Text	<b>OpenGVLab/InternV1.2-8B (best)</b>	40B	55.2	298.7	5391
Text to Image	<i>Kwai-Kolors/Kolors (efficient)</i>	3B	1056.4	2625.5	1914
Text to Image	<b>playgroundai/playground-v2.5-1024px-aesthetic (best)</b>	3B	1123.0	3214.5	23332
Text Classification	<i>novasource/nvidia_en_400M_v3 (efficient)</i>	435M	86.7	5824.7	385014
Text Classification	<b>nvidia/NV-Embed-v2 (best)</b>	8B	90.4	12832.5	324552
Translation	<i>google/glm-4-large (efficient)</i>	738M	32.0	111.0	1028285
Translation	<b>google/glm-4-large (best)</b>	11B	32.1	442.0	1648300
Audio Classification	<i>AI-MUHubert-base-audiovec (efficient)</i>	94M	55.0	388.6	146
Audio Classification	<b>AI-MUHubert-large-audiovec (best)</b>	315M	58.3	766.2	98
Image Segmentation	<i>BDEA-Research/grounding-dino-base (efficient)</i>	223M	60.8	68.2	1064357
Image Segmentation	<b>OpenGVLab/InternImage_V2_2kto1k_640 (best)</b>	1B	62.9	175.2	23
Time Series Forecasting	<i>llm-granite/granite-timeseries-patch8 (efficient)</i>	616M	0.6	405.2	7030
Time Series Forecasting	<b>Salesforce/mobai-1.8-B-small (best)</b>	14M	0.6	5606.5	56095
Code Generation	<i>Qwen/CodeQwen1.5-7B Chat (efficient)</i>	7B	55.1	7518.2	60861
Code Generation	<b>m-a-p/OpenCodeInterpreter-478-33B (best)</b>	33B	55.8	21035.5*	507
Mathematical Reasoning	<i>novasource/Mistral-7B-Instruct-v0.1 (efficient)</i>	7B	18.2	7688.2*	280741
Mathematical Reasoning	<b>Qwen/Qwen-14B-Chat (best)</b>	14B	22.3	12004.3*	257
Text Clustering	<i>novasource/nvidia_en_400M_v3 (efficient)</i>	435M	56.7	5824.7	385014
Text Clustering	<b>nvidia/NV-Embed-v2 (best)</b>	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**.

performing models which are not yet massively used by the community. *Text Classification* (see Figure 5c) 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 of 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 *Kolors* 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.

#### A. AI Inference Energy Consumption Measurement Methodology

Precise measurements of the energy consumption are crucial for selecting energy-efficient models during inference. Several works [42], [43], [44], [45] have proposed the energy monitoring of AI models in order to identify opportunities for enhancing energy efficiency. These works usually use specialized software-based tools for measuring the models energy

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 of 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.

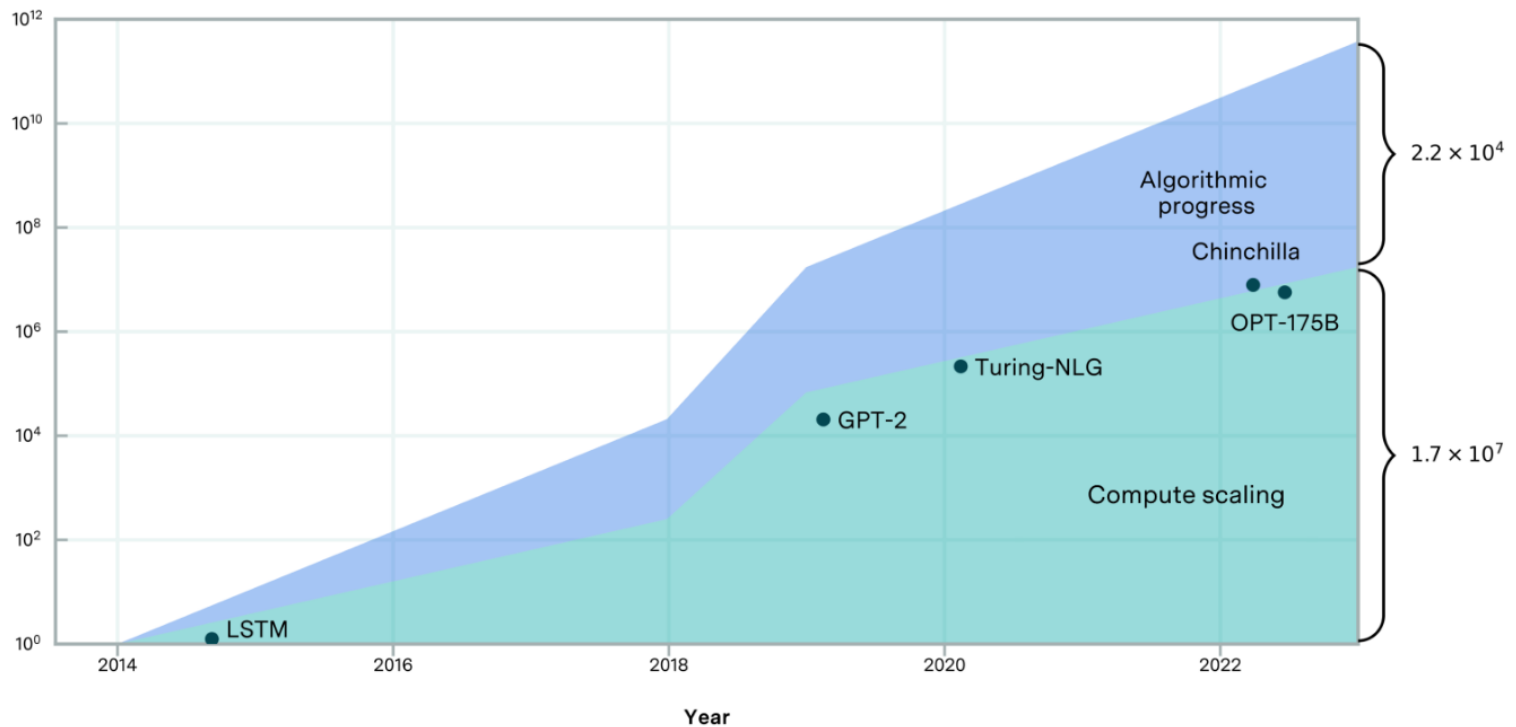
# Transformer Architecture

Full context -- Massive scale-up of parallel computation

Relative contribution of compute scaling and algorithmic progress

EPOCH AI

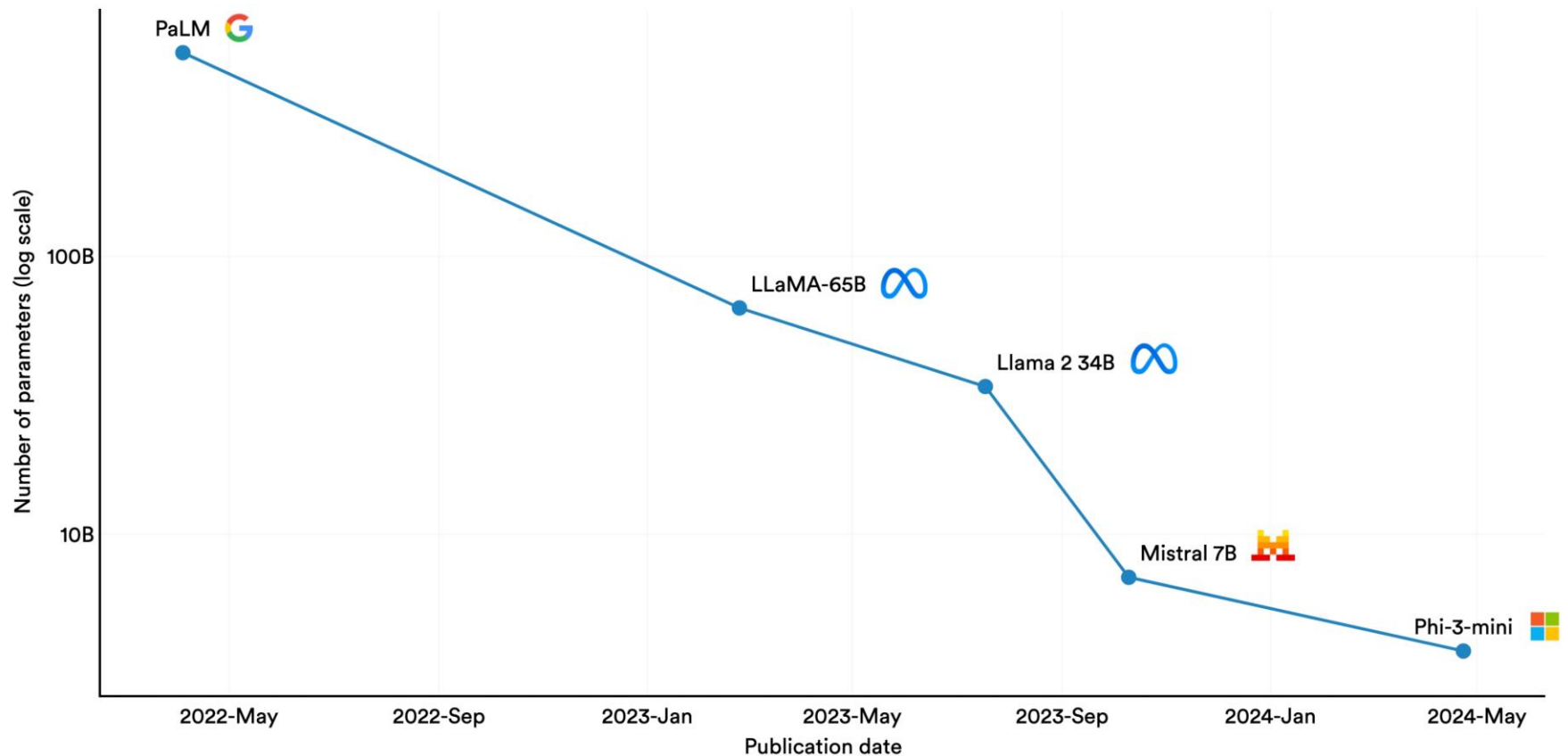
Effective compute (Relative to 2014)



# Not just Bigger and Better

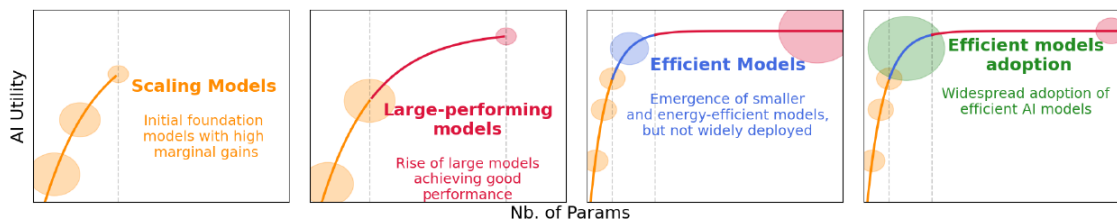
## Smallest AI models scoring above 60% on MMLU, 2022–24

Source: Abdin et al., 2024 | Chart: 2025 AI Index report

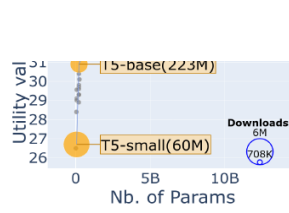
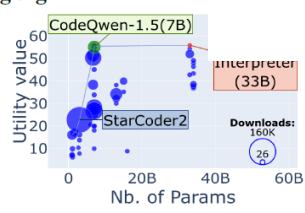
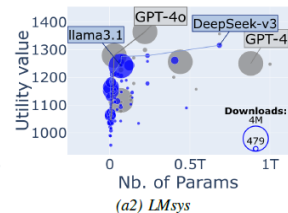
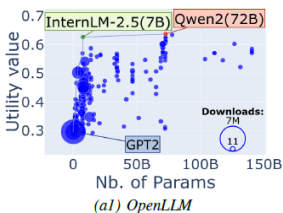




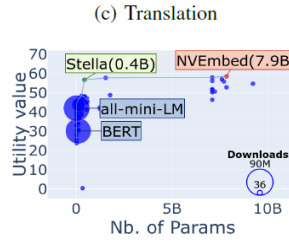
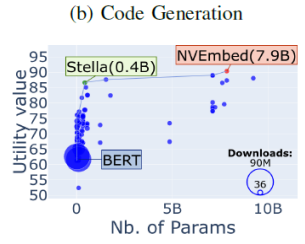
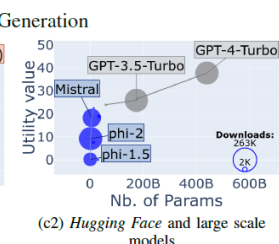
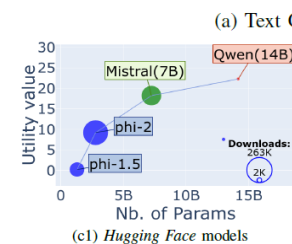
# Small is Sufficient



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



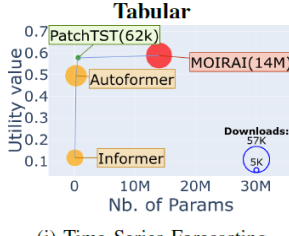
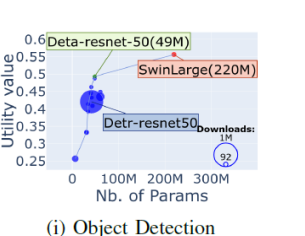
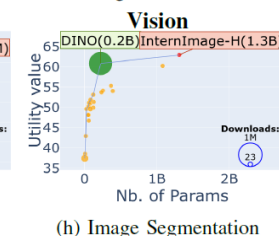
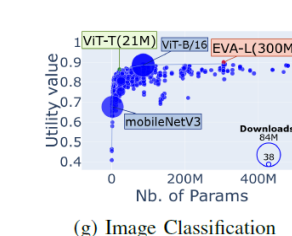
Nb. of Params



(d) Mathematical Reasoning

(e) Text Classification

(f) Text Clustering

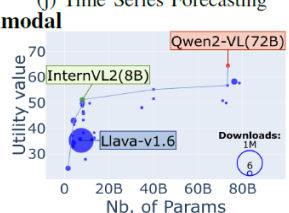
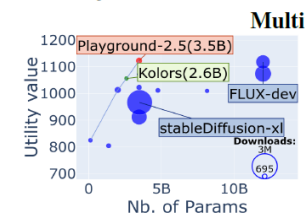
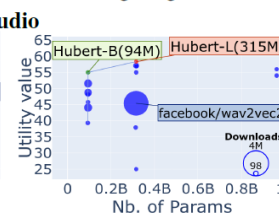
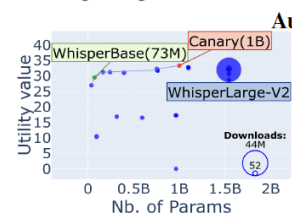


(g) Image Classification

(h) Image Segmentation

(i) Object Detection

(j) Time Series Forecasting



(k) Speech Recognition

(l) Audio Classification

(m) Text to Image

(n) Image-Text to Text

# 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

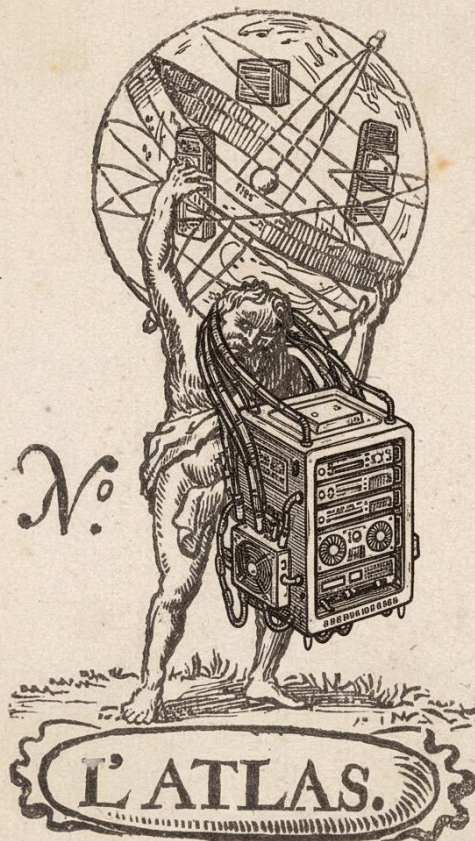
Concerns about privacy incl. inconsistent filtering of sensitive information

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