Bigger, Better, and Shinier:

A Technical and Ecosystem view of Advanced AI, And the misdirection of “frontier models”

Dr. Yacine Jernite, Head of ML & Society, Hugging Face

@ NYU-KAIST Summit on Building Governance Infrastructure for Frontier AI, 06/02/2026

Hugging Face

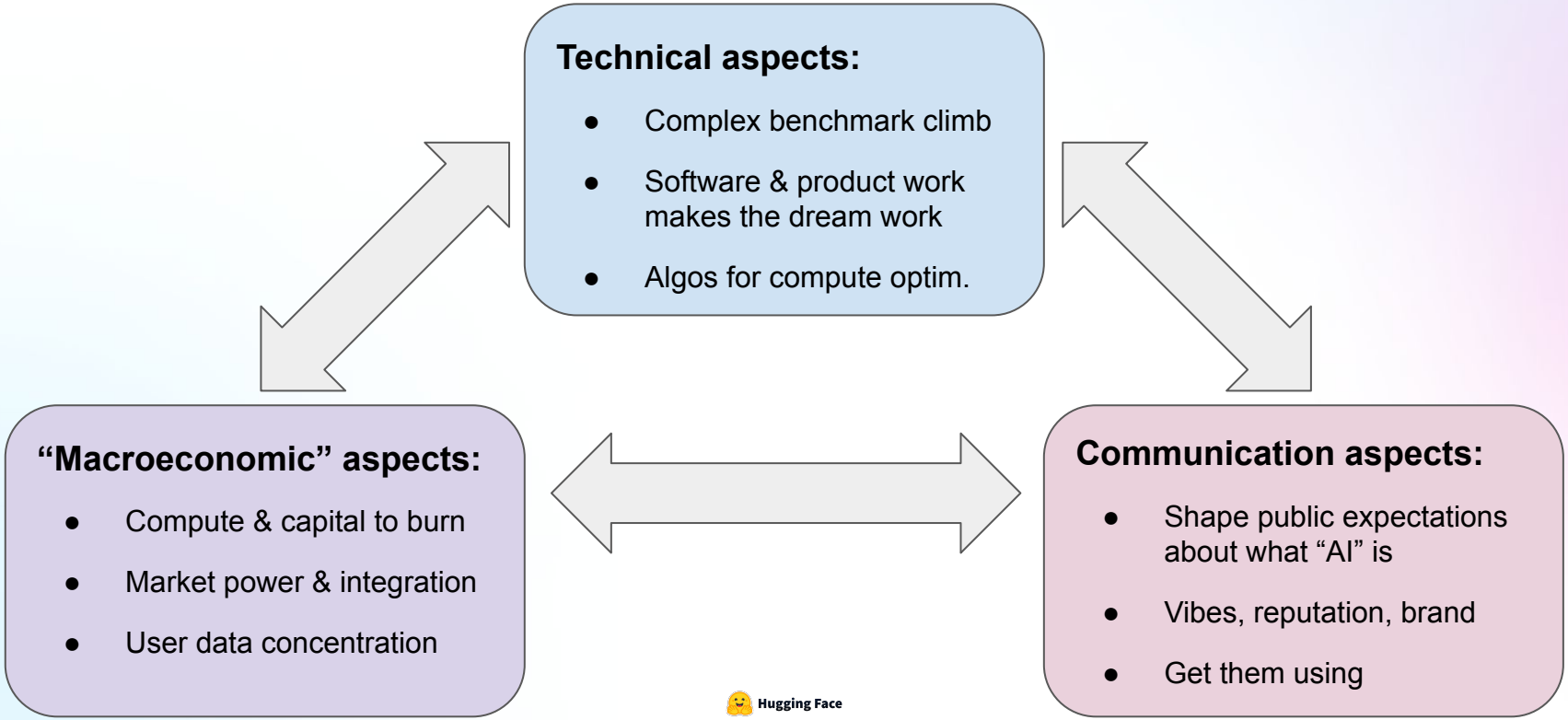
“Frontier” AI in 2026

- Why the “frontier” framing? *Who does it benefit?*
  - Top benchmark scores matter for diffusion, but << adoption & integration
  - Implicature of singular frontier: bigger-is-better and centralization
  - Some other “frontier**s**” in AI
    - 💸 Training efficiency: Alibaba Qwen-(Coder)-Next, base model <1.5M$
    - 🔎 Transparency & Governance
      - 🎓 SmolLM, OLMo, Marin | 🧑‍💼 IBM Granite, NVIDIA Nemotron
    - 🧬 Domain: AlphaFold, OpenFold, Evo2
    - 🐜 Size: Weibo-VibeThinker, edge Claude 4-level math+code
    - 🏃 Speed: OpenAI GPT-OSS, FAST and furious
    - 📜 OCR/input data: Zhipu GLM-OCR-0.5B, speed, custom and performance

- For this talk:
  - Advanced AI, flagship models, very large models…
  - View beyond “out-of-nowhere” aggregated benchmark scores
  - Acknowledge the conditions that create Advanced AI

“Frontier” AI in 2026:

Technical aspects and broader context



Let’s start with the “technical”:

Technical aspects:

- Complex benchmark climb
- Software & product work makes the dream work
- Algos for compute optim.

Some Big Ideas of 2016-2026

From Pre-training to Foundation Models

BERT, transformers, CRFM, etc.

Scale Is All You Need

Especially when you have all the compute, BUT GPT-4.5

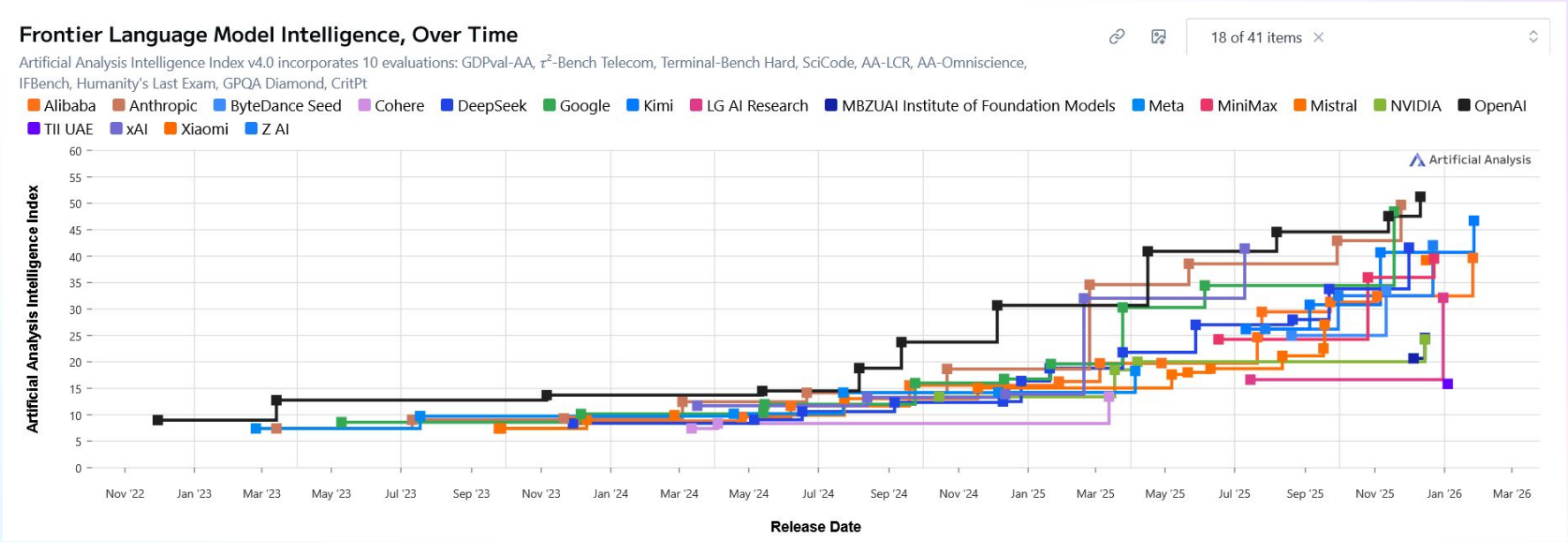
Synthetic Data Works Too

Reinforcement learning, RLVR, Data Augmentation, Data Selection++

Scale is All You Need Continued - Inference-Time

Chain-of-thought, “reasoning”, or guided output flops

Benchmark Climb in Action



Plot sourced from: https://artificialanalysis.ai/

Benchmark Climb in Action

- Looks like a simple narrative:
  - Top spots alternating between OpenAI - Anthropic - Google -xAI
  - Larger and larger models (presumably)
- But very strong caveats:
  - How representatives are those benchmarks? *And of what?*
  - What's missing from this story?
  - Is Scale All there Is?

“Evaluation is Broken”... from the start

Challenges of evaluation science

Replicability, hardware-bound systems, data contamination

From research to commercial systems

Closed-door evaluation under unknown variable conditions, with selection bias, apple-to-orange

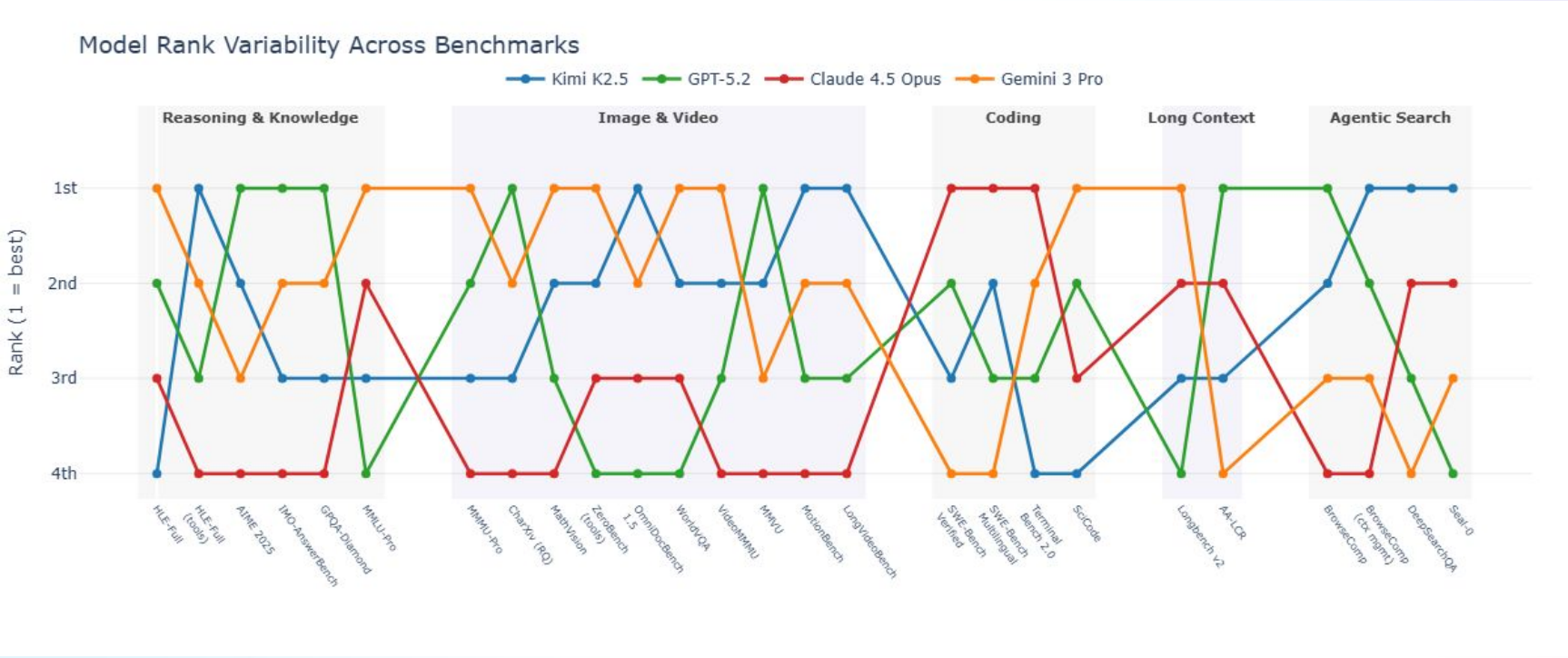
What’s “intelligence”?

Construct validity, anthropomorphic assumptions about “capabilities”

Which “one to rule them all”?

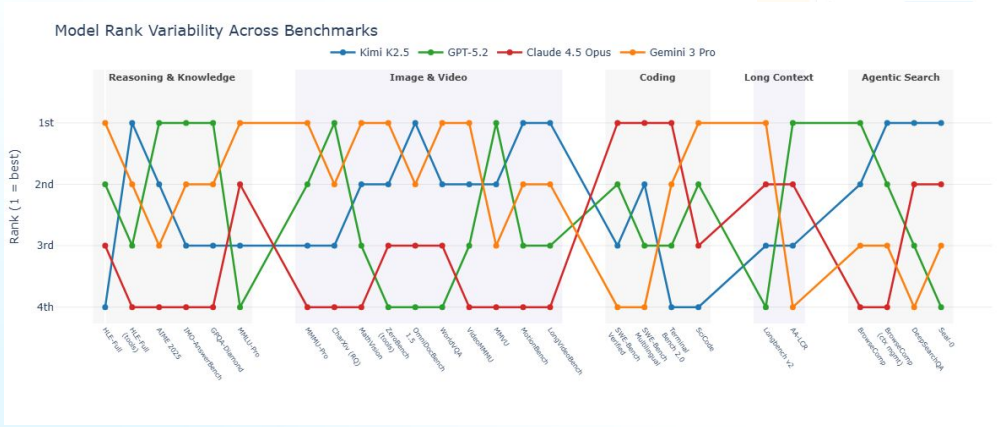
And is that the right question?

Benchmarks in practice: no unified “capabilities”



Benchmark numbers sourced from: https://huggingface.co/moonshotai/Kimi-K2.5

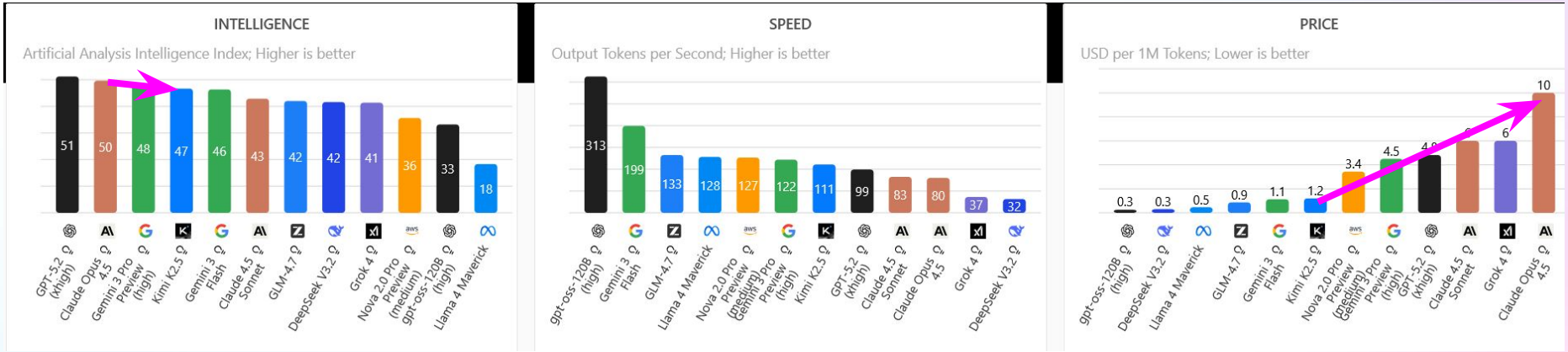
Benchmarks in practice: no unified “capabilities”



[Community Evals: Because we're done trusting black-box leaderboards over the community]

Benchmark numbers sourced from: https://huggingface.co/moonshotai/Kimi-K2.5

Benchmarks in practice: the cost dimension

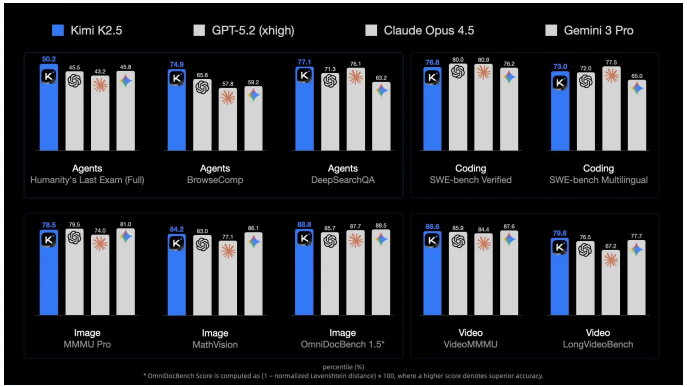


How well does “unlimited inference budget” match governance questions?

Plot sourced from: https://artificialanalysis.ai/

Benchmarks in practice: the forest for the tree

- From the “frontier” perspective, only US companies through 2024
- January 2025: “DeepSeek moment” - 6M$ will get you close!
- July 2025: Moonshot Kimi-K2, the “frontier” is Chinese
- February 2026, flurry of releases pre-New Year



Scientific Tasks

Benchmark	Description	Intern-S1-Pro 1T-A22B	Qwen3-VL-235B-Thinking 235B-A22B	Kimi-K2.5 1T-A32B	GPT-5.2 –	Gemini-3-Pro –
SciReasoner	Scientific Reasoning	**55.5**	11.9	15.3	13.6	14.7
SFE	Scientific Multimodal Tasks	52.7	41.4	53.7	47.5	**58.9**
SmolInstruct	Small Molecule	**74.8**	36.6	53.5	48.2	58.3
MatBench	Materials Property Prediction	**72.8**	49.7	60.0	53.6	64.9
Mol-Instructions	Bio-molecular Instruction	**48.8**	8.9	20.0	12.3	34.6
MicroVQA	Biological Microscopy	63.3	53.8	55.4	60.4	**69.0**
Biology-Instruction	Multi-Omics Sequence	**52.5**	6.2	10.7	10.2	12.0
XLRS-Bench	Remote Sensing	**52.8**	51.2	46.4	50.4	51.8
MSEarth-MCQ	Earth Science	65.2	52.7	61.9	62.6	**65.8**

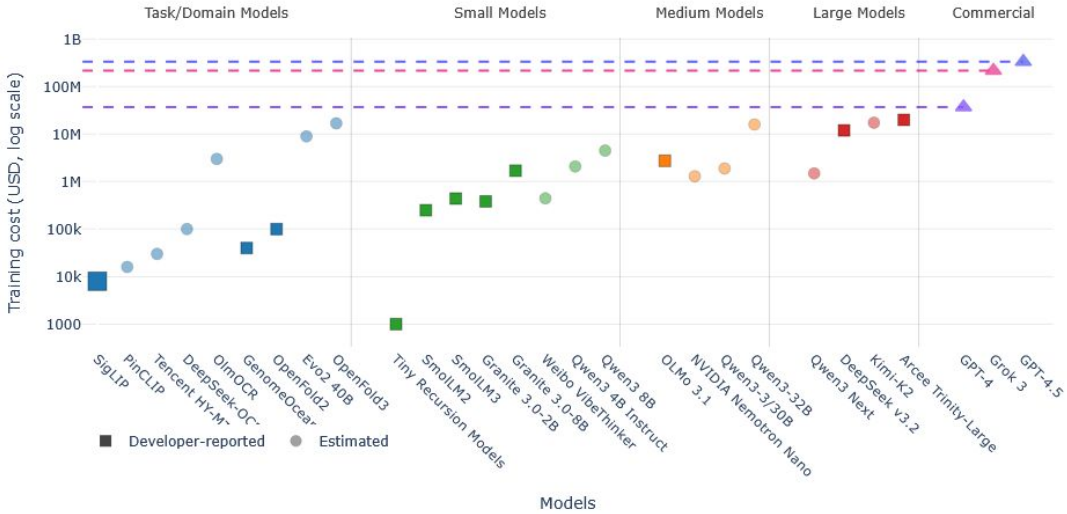
General Tasks

Benchmark	Description	Intern-S1-Pro 1T-A22B	Qwen3-VL-235B-Thinking 235B-A22B	Kimi-K2.5 1T-A32B	GPT-5.2 –	Gemini-3-Pro –
MMMU-Pro	Knowledge & Reasoning	72.8	69.9	78.5	79.5	**81.0**
MMLU-Pro	Knowledge & Reasoning	86.6	83.4	87.1	85.9	**89.3**
AIME-2025	Math Reasoning	93.1	90.0	96.1	**100.0**	95.0
IMO-Answer-Bench	Math Reasoning	77.3	72.3	81.8	**86.3**	81.3
RefCOCO-avg	Visual Grounding	**91.9**	91.1	87.8	54.9	76.2
IFBench	Instruction Following	71.2	58.7	69.7	**75.4**	70.4
OCRBench V2 (ENG / CHN)	OCR	60.1 / 60.6	66.8 / **63.8**	64.2 / 57.4	56.4 / 54.6	**68.0** / 52.5
SArena (Icon)	SVG Generation	**83.5**	76.3	77.3	55.5	82.6
LCB V6	Code	74.3	72.0	85.0	**87.7**	86.9
GAIA (Text-Only)	Agent	77.4	47.8	**79.9**	71.1	75.5
Tau²-Bench	Agent	80.9	57.4	76.8	76.6	**85.4**
ScreenSpot V2	Agent & Grounding	93.6	92.8	92.4	49.4	**94.7**

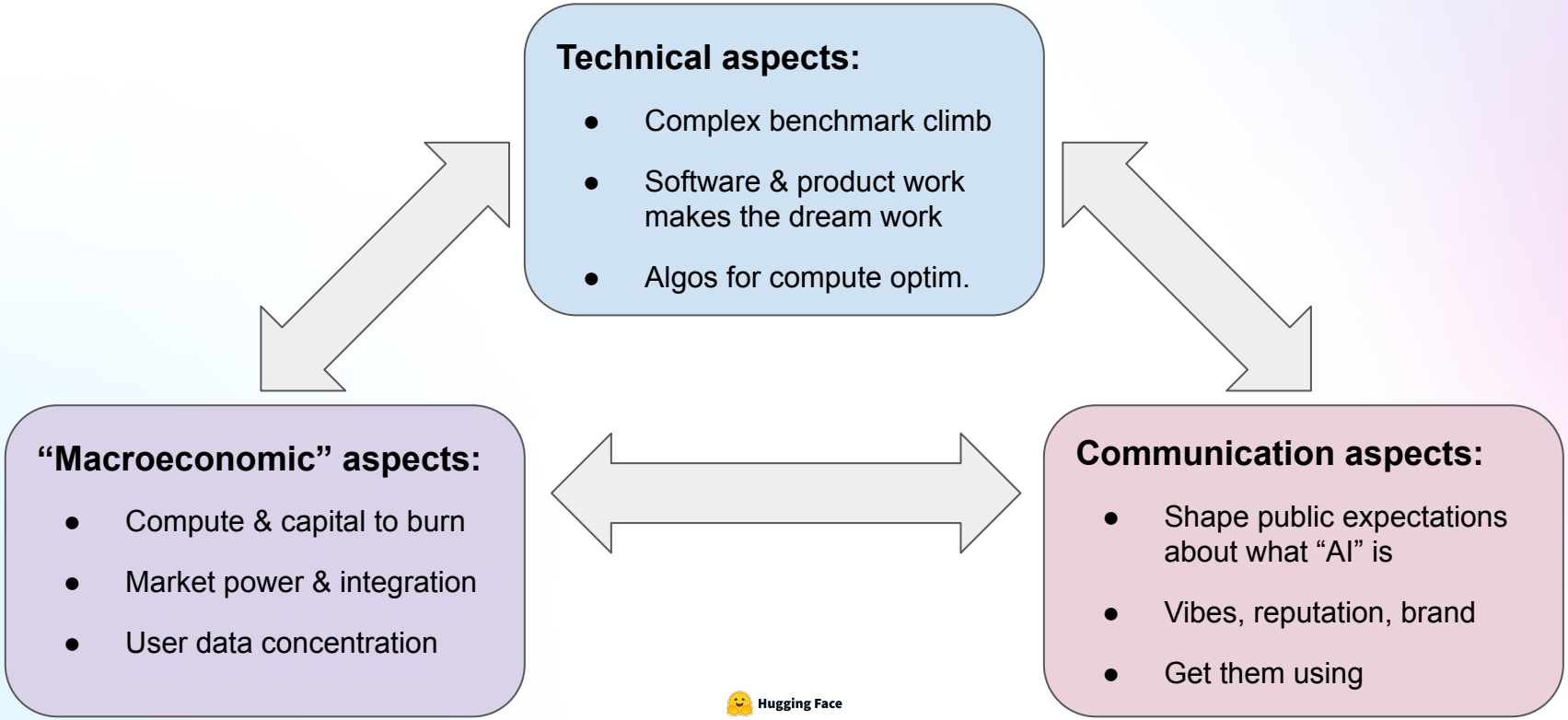
Benchmarks in practice: the Smol “frontier”

- For individual benchmarks, device-side models are also catching up:
  - Weibo VibeThinker - 1.5B, Claude-4 level on math/coding
  - GLM-4.7Flash - 3/30B, beats 8/2025 SOTA on LCBv6
  - Tencent Hunyuan-MT - 1.5B SOTA on machine translation

- Training costs to match:



So what are the benchmarks missing?

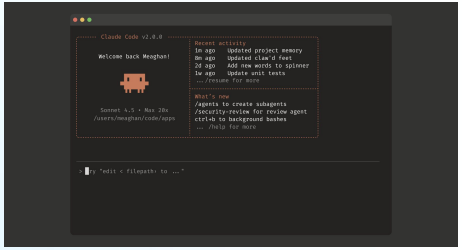


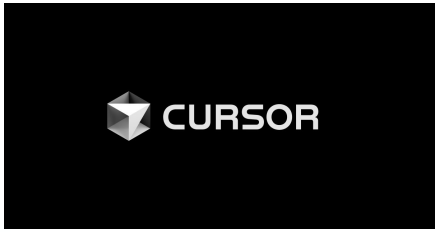
So what are the benchmarks missing?

- Nagle & Yue, 2025: The Latent Role of Open Models in the AI Economy
  - For equal benchmarks, people are spending >+25B$ on proprietary models, why?
  - Competition questions and brand trust, but also likely broader use case coverage
  - Benchmark datasets cover particular use cases, US AI companies collect data about all others!
  - Also helps account for 10x training cost on US models

The (Still) Understated Importance of Data

- From GPT3 to ChatGPT: how are users querying the model?
- Software/UI work makes the dream work:
  - OpenHands: framework == model for agent performance
  - Claude Code: from beloved UI to top coding model
  - Cursor IDE: from wrapper to own model





- Overall recipe: get users using, get model training!

The (Still) Understated Importance of Data

- Overall recipe: get users using, get model training!
- But what does it mean for [privacy|safety|competition]?
  - Incentive to maximize data concentration
    - User Privacy and Large Language Models: An Analysis of Frontier Developers' Privacy Policies
    - Risky Business: Advanced AI Companies' Race for Revenue - Center for Democracy and Technology

The (Still) Understated Importance of Data

- Overall recipe: get users using, get model training!
- But what does it mean for [privacy|safety|competition]?
  - Incentive to maximize data concentration
  - Incentive to push users to explore less safe/appropriate settings
    - OpenAI's Sam Altman touts benefit of GPT-5 for healthcare (followed by ChatGPT Health Announcement)
    - Coding Agents / Laptop assistants shipped with extensive default access

The (Still) Understated Importance of Data

- Overall recipe: get users using, get model training!
- But what does it mean for [privacy|safety|competition]?
  - Incentive to maximize data concentration
  - Incentive to push users to explore less safe/appropriate settings
  - Incentive to systematically overrepresent system capabilities - and threat!
    - In Weak Job Market, Middle Managers Increasingly Forced to Feign AI Success | TechPolicy.Press
    - FTC Announces Crackdown on Deceptive AI Claims and Schemes | Federal Trade Commission
    - Lessons from a Chimp: AI "Scheming" and the Quest for Ape Language

The (Still) Understated Importance of Data

- Overall recipe: get users using, get model training!
- But what does it mean for [privacy|safety|competition]?
  - Incentive to maximize data concentration
  - Incentive to push users to explore less safe/appropriate settings
  - Incentive to overrepresent system capabilities, and threat!
  - Data hoarding practices are often destructive:
    - Harm to digital platform health
    - Harm to journalism revenue streams
    - Harm to community software contributions
    - …

TLDR; Frontier Model vs Data Concentration

- The last 5 years have seen an undeniable trend of larger and more expensive models pushing more complex benchmarks
- At the benchmark level; much cheaper (open) models have gone from following to catching up
- The “frontier” remains for who can collect the most use data - and the compute capital to use it in training. New use cases start with data/deployment.
- To govern AI models, look to the data, compute components, and narratives being crafted 🤗

Thank you for your attention!

Questions?