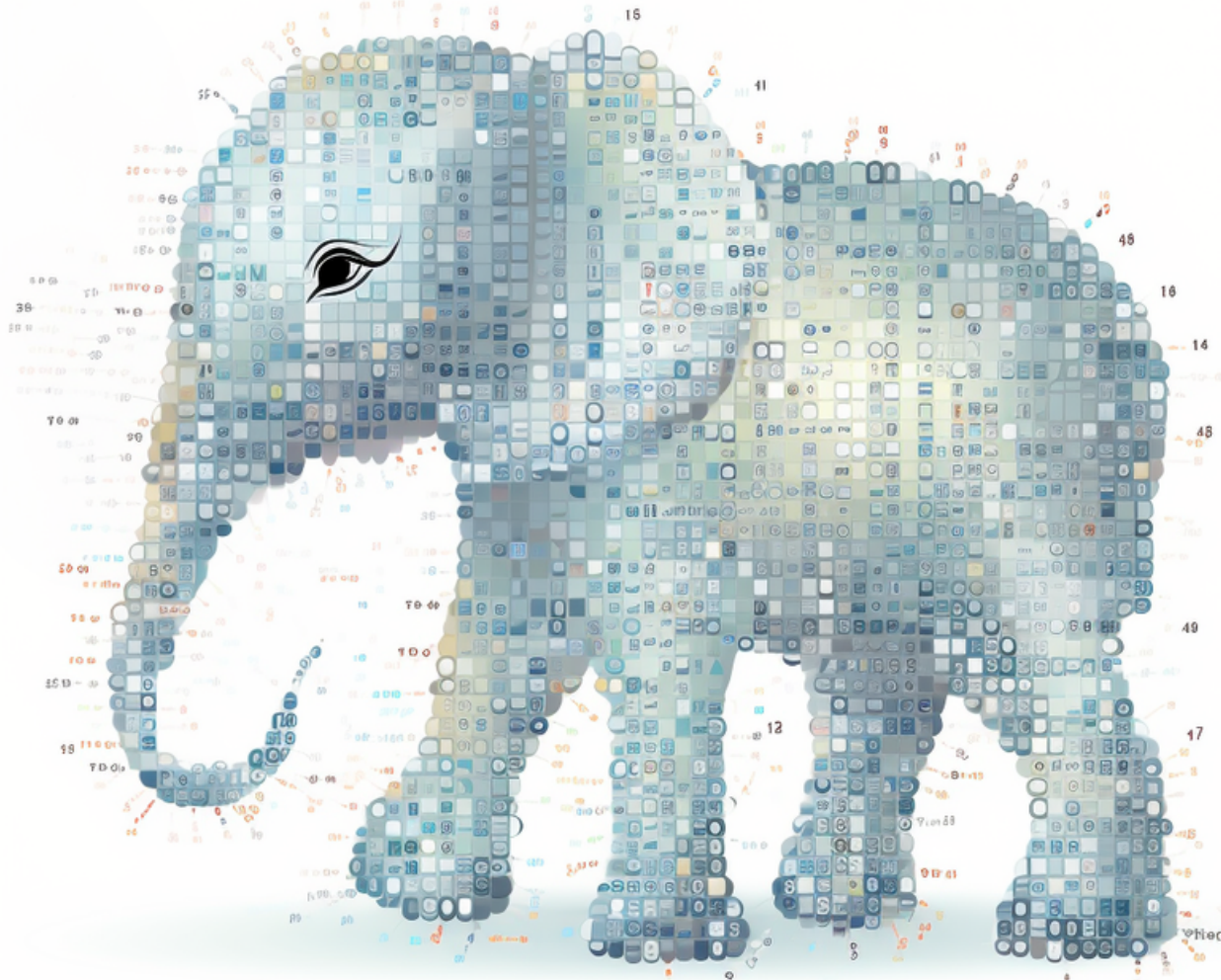


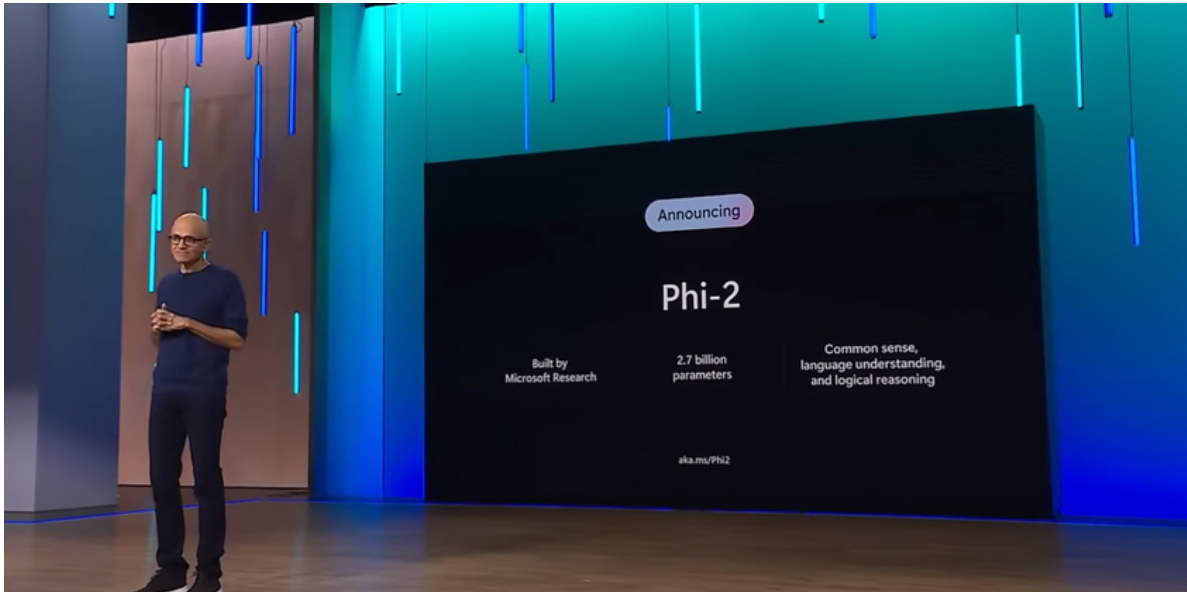
HQQ: Half-Quadratic Quantization of Large Machine Learning Models

- من أكبر المشاكل لي كيلقا أي واحد باغي ي déployer شي LLM هي الحجم ديال هاد models لي كيخليهم كيتقامو غالين en terme de mémoire، داكشي علاش كانستعملو quantization/compression لي بطريقة ذكية كتلخص model weights فشي data type لي كيشد بلاصة صغيرة في mémoire بحال int4.
- غالبا هاد quantization كاتخلق لينا واحد l’erreur فالحسابات ديال model ديالنا، داكشي علاش كيخص نديرو calibration لي كاتصح شوية فدوك quantized weights، أو هادشي لي كيتدار في quantization schemes بحال GPTQ & AWQ، ولكن هاد المرحلة كاتحتاج بزاف ديال الوقت أو compute.
- الطريقة لي جات بيها HQQ، أو لي طورها المغربي هشام بدري، كاتستاند على théorie de l’optimisation، أو ماكاتحتاجش calibration حيث كاتعيد صياغة المشكل بحيث كيولي problème de minimisation لي نقدرو نحلوه باستعمال half-quadratic solver فأقل من دقيقة.



- Blog: [mobiusml.github.io/hqq\\_blog](https://mobiusml.github.io/hqq_blog).
- Code: [github.com/mobiusml/hqq](https://github.com/mobiusml/hqq).

Phi-2: The Surprising Power of Small Language Models



6th grade math exam (2023)	Phi-2 (math FT)	GPT 3.5	Bard	Llama-2 70B	Average 6-th grade Student
Translated French exam (National)	17 / 18	15 / 18	15 / 18	11 / 18	8.7 / 18
US exam (NY)	19.66 / 21	18.33/21	17.66/21	14 / 21	10.5 / 21

phi-2  
(2.7B)

A skier slides down a frictionless slope of height 40m and length 80m. What's the skier's speed at the bottom?

The potential energy of the skier at the top of the slope is given by  $mgh$ , where  $m$  is the mass of the skier,  $g$  is the acceleration due to gravity, and  $h$  is the height of the slope. Since the slope is frictionless, all of the potential energy is converted to kinetic energy at the bottom of the slope. The kinetic energy of the skier is given by  $(1/2)mv^2$ , where  $v$  is the speed of the skier at the bottom of the slope. Equating the potential energy and kinetic energy, we get:  $mgh = (1/2)mv^2$ .

Simplifying, we get:  $v^2 = 2gh$

Substituting the given values, we get:

$v^2 = 2(9.8)(40)$

$v^2 = 784$

$v = 28 \text{ m/s}$

Therefore, the skier's speed at the bottom of the slope is 28 m/s.

- Phi-1 هو واحد Small Language Model لي خرجاتو Microsoft في صيف ديال 2023 أو لي بينات فيه بلي model لي صغير نسبيا (1.3B params) يقدر يوصل performance أولا يتفوق على وحدين كبر منو بزاف إلا وفرنا ليه data نقية أو لي سماوها textbook quality جاين منها من internet أو نص مصيبينو باستعمال ChatGPT-3.5.
- التفوق ديال Phi-1 كان باين خصوصا في tasks لي كانت عندهم علاقة ب reasoning & common sense بحال البرمجة و الرياضيات، أو هادشي لي خلاهم يزيديو يبحثو فهاد direction باش يخرجو Phi-1.5 لي زاد عزز نظرية أن Textbook Is All You Need حيث لاحظو بلي فاش نقصو من internet data، نقصو المشاكل لي كان كان كيغاني منهم Phi-1 بحال hallucination & toxicity.
- فشهر décembre زادونا Phi-2 لي كبر من لولين (2.7B params) ولي ماحتاجوش يعاودو entrainment ديالو من الأول، بل استاندو على weights ديال Phi-1.5 باش يديرو knowledge distillation أو يكملو training على data جديدة.
- فاش دربوه على تمارين ديال الرياضيات (1M)، Phi-2 قدر يجيب 17/18 في الامتحان الوطني الفرنسي أو 19,66.21 فالأمريكي !
- وأخيرا هاد الشهر Phi-2 ولا Open Source فاش Microsoft بدلات licence ديالو لي ولات MIT، يعني community غادي تبدا تجرب فيه بكل أريحية.

- Blog: [microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models](https://microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models)
- Model: [huggingface.co/microsoft/phi-2](https://huggingface.co/microsoft/phi-2)

Rabbit R1: The AI Pocket Agent

- هاد الشهر فمعرض CES العالمي لي تقام في Las Vegas، لاحظنا بلي AI كانت المحور ديال بزاف products، أنجح واحد فيهم هو Rabbit R1 لي كيشبه لشئ téléphone ولكن فالحقيقة هدا gadget لي لهدف ديالو هي أنه يخليك تنقص من الاستعمال ديال téléphone، كيفاش ؟
- فيه Operating System سميتو Rabbit OS مبني على واحد النوع جديد ديال Foundation Models سماوه Large Action Model. هاد device كاتقدر تهدر معاه بصوتك فأي لقطة أو تقول ليه أش بغيتيه يدير، مثلا يعيط لك على Uber لشئ بلاصة أولا يشير لك شي حاجة من Amazon، أو هو كيتكلف أو كيدير الواجب باستعمال models لي intégrés فيه ولي داخل فيم LLM, LAM, Text To Speech, Speech To Text, Vision Model، إلى آخره.

- Keynote: [rabbit.tech/keynote](https://rabbit.tech/keynote)



official partner



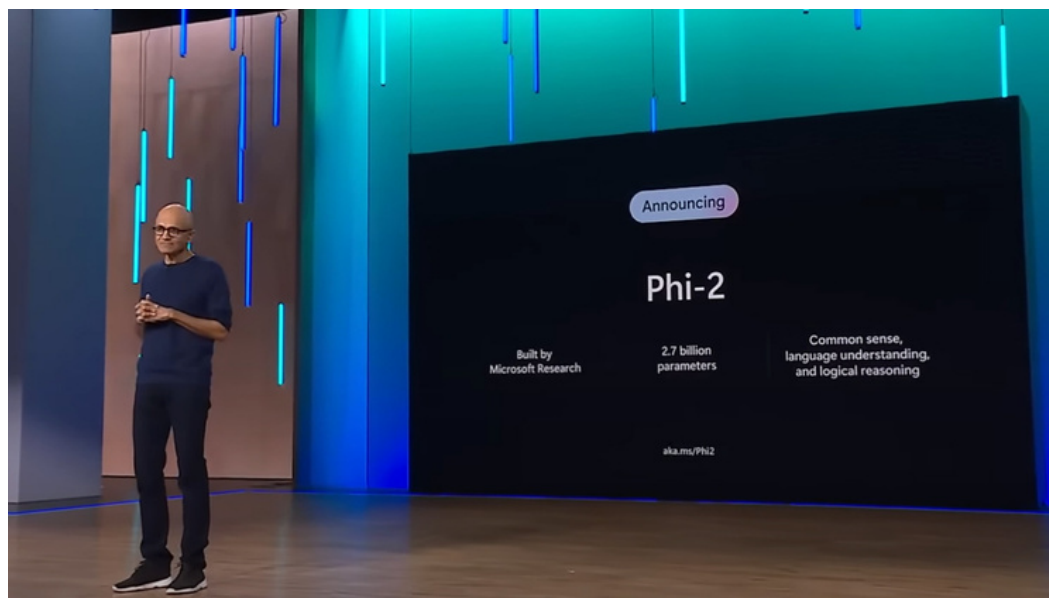


## HQQ: Half-Quadratic Quantization of Large Machine Learning Models


- One of the major challenges encountered when deploying Large Language Models (LLMs) is the considerable size of these models, rendering them resource-intensive in terms of memory. For that reason, using quantization/compression techniques is deemed necessary to intelligently summarize the model weights into a data type taking smaller memory space, such as int4.
- Quantization often introduces errors in the calculations of our model. Consequently, a calibration process is undertaken to rectify these discrepancies in the quantized weights. This calibration step is pivotal and is addressed in quantization schemes like GPTQ and AWQ. However, this phase demands a significant amount of time and computational resources.
- Half-Quadratic Quantization (HQQ) method, developed by Moroccan researcher Hicham Badri, is an innovative approach leveraging optimization theory and eliminates the need for calibration. This reformulation to a minimization problem enables efficient resolution using half-quadratic solver in less than a minute.
  - **Blog:** [mobiusml.github.io/hqq\\_blog](https://mobiusml.github.io/hqq_blog).
  - **Code:** [github.com/mobiusml/hqq](https://github.com/mobiusml/hqq).



## Phi-2: The Surprising Power of Small Language Models



6th grade math exam (2023)	Phi-2 (math FT)	GPT 3.5	Bard	Llama-2 70B	Average 6-th grade Student
Translated French exam (National)	17 / 18	15 / 18	15 / 18	11 / 18	8.7 / 18
US exam (NY)	19.66 / 21	18.33/21	17.66/21	14 / 21	10.5 / 21



phi-2  
(2.7B)

A skier slides down a frictionless slope of height 40m and length 80m. What's the skier's speed at the bottom?

The potential energy of the skier at the top of the slope is given by  $mgh$ , where  $m$  is the mass of the skier,  $g$  is the acceleration due to gravity, and  $h$  is the height of the slope. Since the slope is frictionless, all of the potential energy is converted to kinetic energy at the bottom of the slope. The kinetic energy of the skier is given by  $(1/2)mv^2$ , where  $v$  is the speed of the skier at the bottom of the slope. Equating the potential energy and kinetic energy, we get:  $mgh = (1/2)mv^2$ .

Simplifying, we get:  $v^2 = 2gh$

Substituting the given values, we get:

$v^2 = 2(9.8)(40)$

$v^2 = 784$

$v = 28 \text{ m/s}$

Therefore, the skier's speed at the bottom of the slope is 28 m/s.

- Phi-1, a Small Language Model, was released by Microsoft in the summer of 2023. Despite its relatively modest size (1.3B params), it has shown noteworthy performance, surpassing larger models when provided with high-quality data -referred to as textbook quality-. with half of this data sourced from the internet, and the other half generated using ChatGPT-3.5.
- Phi-1's outmatch was particularly evident in tasks related to reasoning and common sense, such as programming and mathematics. This success motivated them to further explore in this direction, leading to the development of Phi-1.5. This has strengthened the theory that 'Textbook Is All You Need' by noticing that reducing dependence on internet data has addressed challenges faced by Phi-1, like hallucination and toxicity.
- In December, the unveiling of Phi-2, a larger model with 2.7B params, marked a significant advancement. Notably, the model's development skipped training from scratch by leveraging Phi-1.5's weights to perform knowledge distillation and continued training on new data.
- Remarkably, when tested on mathematics exercises (1M), it achieved an impressive score of 17/18 in the French national examination and 19,66/21 in its American counterpart.
- In a final development this month, Phi-2 has become open-source, with Microsoft changing its license to MIT. This shift allows the wider community to experiment and engage with Phi-2 more freely.

- **Blog:**[microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models](https://microsoft.com/en-us/research/blog/phi-2-the-surprising-power-of-small-language-models)
- **Model:** [huggingface.co/microsoft/phi-2](https://huggingface.co/microsoft/phi-2)

## Rabbit R1: The AI Pocket Agent

- This month at the CES international in Las Vegas, AI took center stage with numerous products, and one notable success was the Rabbit R1. The phone-like gadget is designed to reduce your dependence on phone usage. Curious how?
- It runs on an operating system called Rabbit OS, built on a novel type of Foundation Models named Large Action Model. This device allows you to interact with it using your voice for any task. For instance, it can call an Uber for you to a specific location, order something from Amazon, or perform duties using its embedded models such as LLM, LAM, Text To Speech, Speech To Text, Vision Model, and more.

- **Keynote:** [rabbit.tech/keynote](https://rabbit.tech/keynote)



official partner

