

#CONF'

Maxime THOONSEN

CTO @ Theodo / Orga @ Gen AI Paris



Le 23/01/23 à 19h

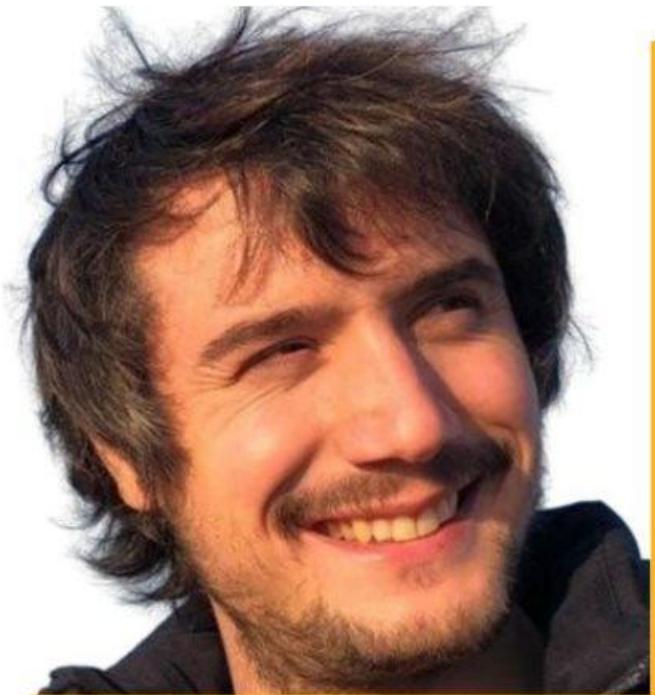


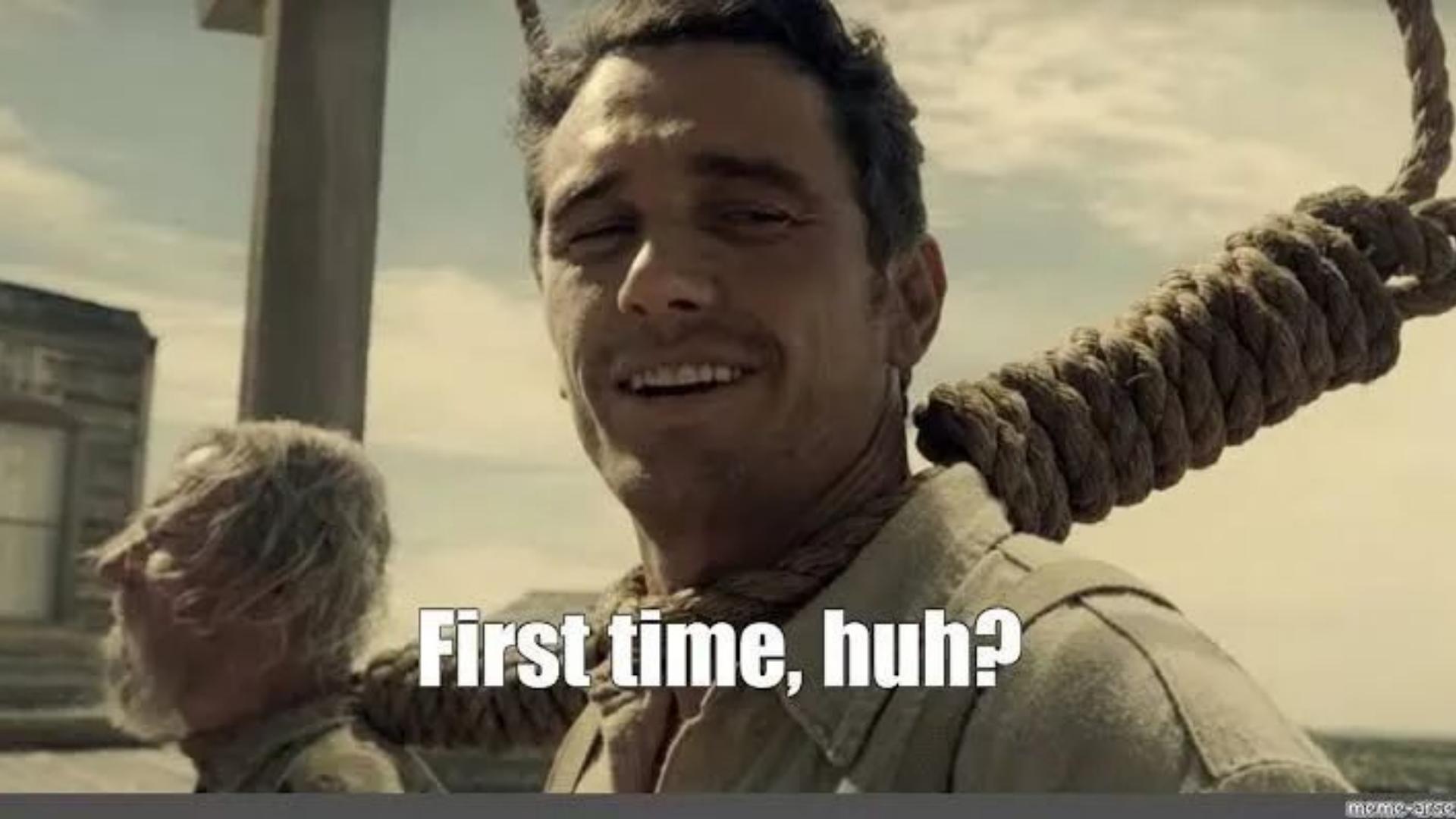
lonestone



[sfēir]

*“REX sur la création d'Auto PHP,
un agent autonome open source”*



A close-up photograph of a man with dark hair and a beard, smiling broadly. He is wearing a light-colored, collared shirt. A thick, dark rope is wrapped around his neck, with one end resting against his shoulder and the other extending towards the top right corner of the frame. In the bottom left corner, the back of another person's head and shoulders are visible, showing short, wavy hair and a greenish-brown jacket.

First time, huh?



Schedule



1- News (20min)



2- Talk (1h30)



3- Enjoy ;)

Sponsor turbo-platinum



ÉDITEUR DE LOGICIELS
POUR LES EXPERTS-COMPTABLES

[sf≡ir]

lonestone



LLM / LVM for mobile

 Privacy

 Cost for developer

 Battery

 Smaller model

 Fast (no network latency)

 Slow (token/s)

Qualcomm



On-device Personalization
Qualcomm® Sensing Hub

Qualcomm® AI Stack Models -
Fully optimized models

Fastest stable diffusion in the world
<1 sec per image

First to support multi-modality gen AI models



◆ Text Prompt. |

15 Tokens/sec

Meta Llama 2 powered AI assistants

"Voice prompt."

/ Photo Prompt /



10Billion+

Parameter support



Google Gemini Nano

2 versions:

- Nano-1: 1.8B parameters
- Nano-2: 3.25B parameters

Distillated from Gemini Ultra
4-bit quantization

	Gemini Ultra (pixel only)	Gemini Pro (pixel only)	Gemini Nano 2 (pixel only)	Gemini Nano 1 (pixel only)	GPT-4V
MMMU (val) Multi-discipline college-level problems (Yue et al., 2023)	59.4% pass@1	47.9%	32.6%	26.3%	56.8%
TextVQA (val) Text reading on natural images (Singh et al., 2019)	62.4% Maj1@32				
DocVQA (test) Document understanding (Mathew et al., 2021)	82.3%	74.6%	65.9%	62.5%	78.0%
ChartQA (test) Chart understanding (Masry et al., 2022)	90.9%	88.1%	74.3%	72.2%	88.4% (pixel only)
InfographicVQA (test) Infographic understanding (Mathew et al., 2022)	80.3%	75.2%	54.5%	51.1%	75.1% (pixel only)
MathVista (testmini) Mathematical reasoning (Lu et al., 2023)	53.0%	45.2%	30.6%	27.3%	49.9%
AI2D (test) Science diagrams (Kembhavi et al., 2016)	79.5%	73.9%	51.0%	37.9%	78.2%
VQAv2 (test-dev) Natural image understanding (Goyal et al., 2017)	77.8%	71.2%	67.5%	62.7%	77.2%



Google Gemini Nano

Usage:

- Text completion
- STT
- Multi-modal ??

Available on Pixel 8 Pro + Samsung S24

Android app target: API level >= 21

Fine-Tuning using LoRA layer

Microsoft Phi-2

2.7 billion parameter
Quantization: FP16

OpenSource: <https://huggingface.co/microsoft/phi-2>

Usage:

- Text completion
- QA
- Code generation 

Microsoft Phi-2

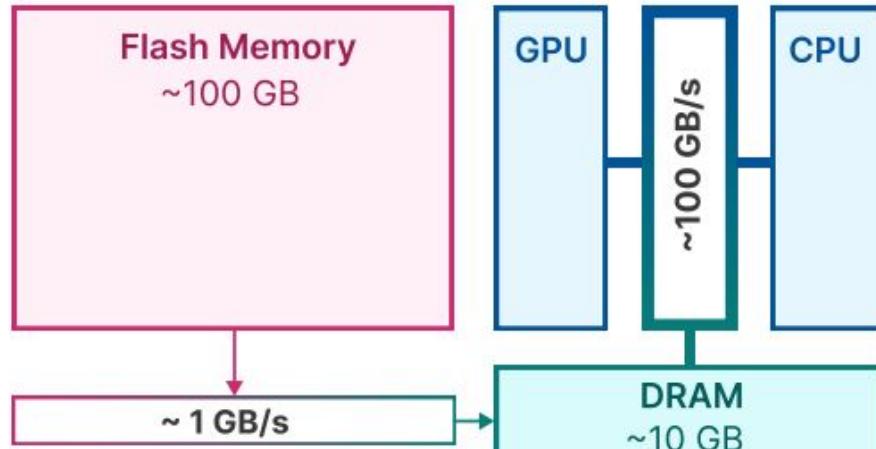
Model	Size	BBH	Commonsense Reasoning	Language Understanding	Math	Coding
Llama-2	7B	40.0	62.2	56.7	16.5	21.0
	13B	47.8	65.0	61.9	34.2	25.4
	70B	66.5	69.2	67.6	64.1	38.3
Mistral	7B	57.2	66.4	63.7	46.4	39.4
Phi-2	2.7B	59.2	68.8	62.0	61.1	53.7

Table 1. Averaged performance on grouped benchmarks compared to popular open-source SLMs.

Model	Size	BBH	BoolQ	MBPP	MMLU
Gemini Nano 2	3.2B	42.4	79.3	27.2	55.8
Phi-2	2.7B	59.3	83.3	59.1	56.7

Table 2. Comparison between Phi-2 and Gemini Nano 2 Model on Gemini's reported benchmarks.

LLM Inference with Limited Memory: offloading



(a) Bandwidth in a unified memory architecture

- * Pull weight from DRAM on-demand
- * Optimize bandwidth, cache and batching
- * Cool for speculative decoding
- * Cool for Mixture of Experts

LLM Inference with Limited Memory: offloading

Table 2: The I/O latency of OPT 6.7B 16 bit on M1 max for different techniques when half the memory is available

Configuration				Performance Metrics			
Hybrid	Predictor	Windowing	Bundling	DRAM (GB)	Flash→DRAM(GB)	Throughput (GB/s)	I/O Latency (ms)
✗	✗	✗	✗	0	13.4 GB	6.10 GB/s	2130 ms
✓	✗	✗	✗	6.7	6.7 GB	6.10 GB/s	1090 ms
✓	✓	✗	✗	4.8	0.9 GB	1.25 GB/s	738 ms
✓	✓	✓	✗	6.5	0.2 GB	1.25 GB/s	164 ms
✓	✓	✓	✓	6.5	0.2 GB	2.25 GB/s	87 ms

LLM Inference with Limited Memory: offloading



https://github.com/lyogavin/Anima/tree/main/air_llm

Conversational LLM available in Europe

✓ GPT4 / ChatGPT

✓ Palm2 / Bard

✓ HuggingChat

✗ Google Gemini -> ✓ via VPN

✗ Claude 2.1 -> ✓ via Devana or Poe

DeepMind: Research powered by LLM

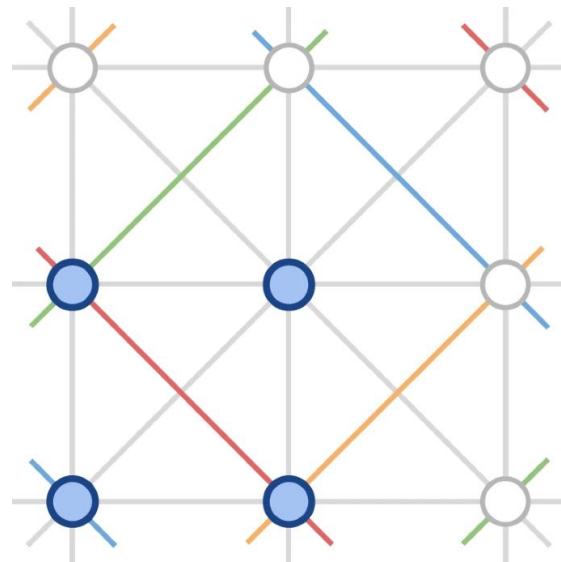
Binpacking



Best-fit heuristic

FunSearch

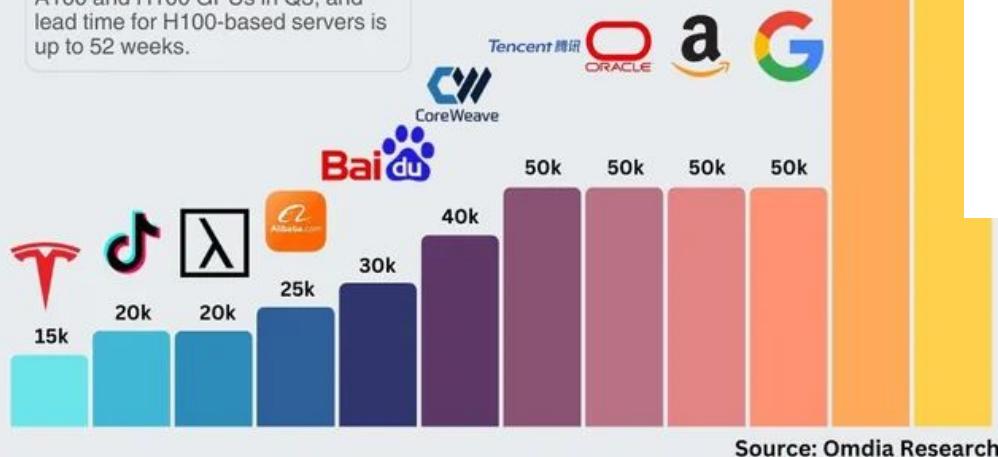
Cap set problem



Nvidia H100 GPU Shipments by Customer

Estimated 2023 H100 shipments by end customer.

Omdia estimates Nvidia sold ~500k A100 and H100 GPUs in Q3, and lead time for H100-based servers is up to 52 weeks.



150k

150k



zuck 🔥

25m ...

We're building a massive amount of infrastructure. At the end of this year, we'll have ~350k Nvidia H100s -- and overall ~600k H100s H100 equivalents of compute if you include other GPUs



Apple ML Ferret

📍 Ferret: Ferret: Refer and Ground Anything Anywhere at Any Granularity

Instructions:

1. Select a 'Referring Input Type'
2. Draw on the image to refer to a region/point.
3. Copy the region id from 'Referring Input Type' to refer to a region in your chat.

ferret_ft_clipL336_vicunaV1-3-13b_3Ep_dataV16_RSamplerV2_checkpoint-final

Image & Sketch

Referring Input Type

Point Box Sketch

Referring Input Cache

[region0] [region1]

Parsed Referring Input

obj0

(region1)

(region0)

FERRET

What's the relationship between object [region0] and object [region1]?

The object [obj0] is a dog, and the object [obj1] is a ferret. The dog and the ferret appear to be sitting together in the snow, suggesting a friendly interaction or companionship.

Output Vis

obj0

obj1

Enter text and press ENTER

Submit

Upvote

Downvote

Regenerate

Clear history

Show location

Midjourney v6

THIS IS
SPARTA!





I'M
YOUR
FATHER

THE
HIS

TM & © 2015 Lucasfilm Ltd. All Rights Reserved.

STAR WARS







Image de film 35mm, plan de deux personnes : un homme noir de 50 ans avec une barbe grise, portant une veste marron et une écharpe rouge, se tenant à côté d'une femme blanche de 20 ans. Elle porte un manteau à carreaux bleu marine et crème et un bonnet tricoté noir. Ils marchent au milieu de la rue à minuit, éclairés par la douce lueur orange des lampadaires.



phone photo of [DECRIRE LA SCENE] posted to reddit in 2019, --style raw --s 0 --ar 9:16 --v 6.0

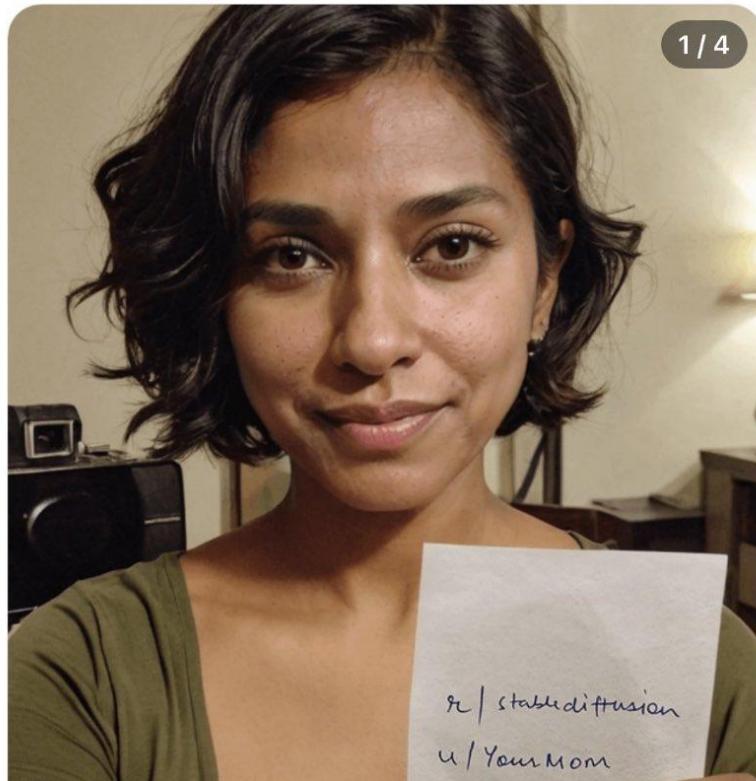


r/StableDiffusion
u/_harsh_ • 13h

Verification post!

Workflow Included

1 / 4



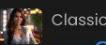
Scénario

Workbench

▼ Concepts



0.75



020

▼ Prompt

close up portrait of a beautiful woman

blonde wearing a red dress

Q Prompt Builder

❖ Prompt Spark

▼ Negative Prompt

blurred faded haze

bad





Search



Train a Model (LoRA)

Blonde Girl (Ghpri)

Model Type

 SDXL LoRA SD 1.5 (legacy)

Start Training

Save As Draft

Advanced Settings

Total Training Steps: 1,750

 Default

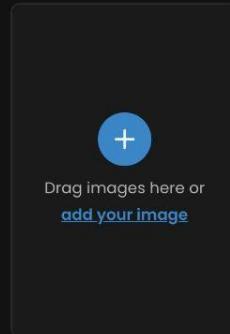
UNet Learning Rate: 6e-5

Text Encoder Training Ratio: 0.20

Text Encoder Learning Rate: 1e-5

Training Images

Delete All



Search

Clear Session

Generate Images

New Model

LoRA Components

New Model

Prompt

Negative Prompt

Reference Image

Select from Library

Sketch

Drop image or import your file

Generative Enhance

Settings

Images: 4

?

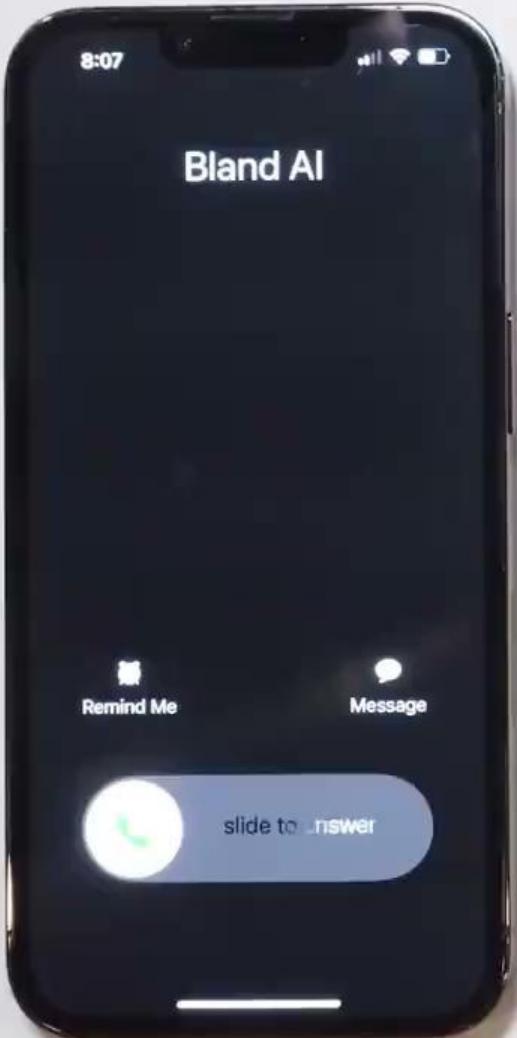
E

Clear Session

The interface is a digital tool for generating images. On the left, there's a sidebar with icons for home, library, and settings. The main area has sections for 'Generate Images' and 'Clear Session'. Under 'Generate Images', there are fields for 'New Model', 'LoRA Components', 'Prompt', 'Negative Prompt', and 'Reference Image' (with options for 'Select from Library', 'Sketch', or 'Drop image or import your file'). Below these are 'Generative Enhance' and 'Settings' sections. The right side shows a 4x4 grid of generated images, each featuring a blonde woman in different styles, such as a red hat, a red jacket, a white shirt, and a red dress, set against a backdrop of a Parisian cafe with red awnings.



Bland AI



January 2024

The world's fastest conversational
AI is released.

MultiOn



The next evolution of Software

Auto | Step

What's next?



Wish Happy Birthday
to friends on fb



Make a Tweet
about MultiOn!

Channel 1 News



+1

Galaxy AI

Galaxy AI est là

Bienvenue dans la nouvelle ère de l'Intelligence Artificielle mobile. Avec le Galaxy S24 Ultra entre vos mains, débloquez de nouveaux niveaux de créativité, de productivité et de possibilité - en commençant par l'outil le plus important de votre vie. Votre smartphone.



Entourer pour chercher¹

Découvrir la fonction Entourer pour chercher ↓



Entourer
pour
chercher



Traduction
instantanée



Assistant
Notes



Assistant
Photo

Rabbit



rabbit™

63

600
100



rabbit™

23

200
100



rabbit™

63

600
100



GPTs store

ChatGPT

Explore GPTs

Today

Fixing a road bike

Calculus homework

Reserve a flight

9:41

Explore

Top Picks My GPTs DALL-E Writing

Featured

Curated top picks from this week

AllTrails
Find trails that fit your nature for your next hike, ride, or run.
By alltrails.com

Consensus
Formerly ResearchGPT. Search 200M academic papers from Consensus, get science-based answers, and draft...
By Krog

Code Tutor
Let's code together! I'm Khanmigo Lite, by Khan Academy. I won't write the code for you, but I'll help you work things out. Can you tell me...
By khanacademy.org

Books
Your AI guide in the world of literature and reading.
By Josh Brent N. Villacido

Trending

Most popular GPTs by our community

DesignerGPT
Creates and hosts beautiful websites
By Pietro Schirano · 50.1k

ScholarAI
Your Research Assistant - I'll help you navigate over a corpus of 200M articles, journals, and books
By scholarai.in · 50.1k

https://chat.openai.com/gpts

My GPTs + Create

GPTs

Discover and create custom versions of ChatGPT that combine instructions, extra knowledge, and any combination of skills.

Search Public GPTs

Top Picks DALL-E Writing Research & Analysis Programming Education Productivity Lifestyle

Featured

Curated top picks from this week

AllTrails
Find trails that fit your nature for your next hike, ride, or run.
By alltrails.com

Consensus
Formerly ResearchGPT. Search 200M academic papers from Consensus, get science-based answers, and draft content with...
By Consensus

Code Tutor
Let's code together! I'm Khanmigo Lite, by Khan Academy. I won't write the code for you, but I'll help you work things out. Can you tell me...
By khanacademy.org

Books
Your AI guide in the world of literature and reading.
By Josh Brent N. Villacido

Trending

Most popular GPTs by our community

DesignerGPT
Creates and hosts beautiful websites
By Pietro Schirano · 50.1k

Canva
Effortlessly design anything: presentations, logos, social media posts and more.
By Canva · 67.3k

ScholarAI
Your Research Assistant - I'll help you navigate over a corpus of 200M articles, journals, and books
By scholarai.in · 50.1k

AI PDF
AI PDF GPT (Top PDF GPT), can handle PDF documents up to 2GB PER FILE, allows 1000s of PDF uploads on mydrive.com
By mydrive.com · 209.5k

Procès Open AI x NYT



The New York Times

EXAMPLE 48: POPE IN BREAK FROM DOCTRINE BACKS SAMESEX CIVIL UNIONS

<https://www.nytimes.com/2020/10/21/world/europe/pope-francis-same-sex-civil-unions.html>
Copyright number: TX 8-926-133 Effective Date: 2020-12-14

Prompt (taken from article):

ROME — Pope Francis expressed support for same-sex civil unions in remarks revealed in a

Output from GPT-4:

documentary film that premiered on Wednesday, a significant break from his predecessors that staked out new ground for the church in its recognition of gay people.

The remarks, coming from the leader of the Roman Catholic Church, had the potential to shift debates about the legal status of same-sex couples in nations around the globe and unsettle bishops worried that the unions threaten what the church considers traditional marriage — between one man and one woman.

“What we have to create is a civil union law. That way they are legally covered,” Francis said in the documentary, “Francesco,” which premiered at the Rome Film Festival, reiterating his view that gay people are children of God. “I stood up for

Actual text from NYTimes:

documentary film that premiered on Wednesday, a significant break from his predecessors that staked out new ground for the church in its recognition of gay people.

The remarks, coming from the leader of the Roman Catholic Church, had the potential to shift debates about the legal status of same-sex couples in nations around the globe and unsettle bishops worried that the unions threaten what the church considers traditional marriage — between one man and one woman.

“What we have to create is a civil union law. That way they are legally covered,” Francis said in the documentary, “Francesco,” which debuted at the Rome Film Festival, reiterating his view that gay people are children of God. “I stood up for

Soirée Meetup

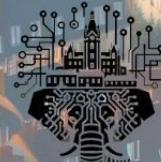
Generative AI par la pratique :

cas concrets d'utilisation d'un LLM en
Java, avec l'API PaLM, Gemini et
LangChain4J



Guillaume LAFORGE

@Google



15 février



19h



e-artsup

#CONF'

Maxime THOONSEN

CTO @ Theodo / Orga @ Gen AI Paris



Le 23/01/23 à 19h

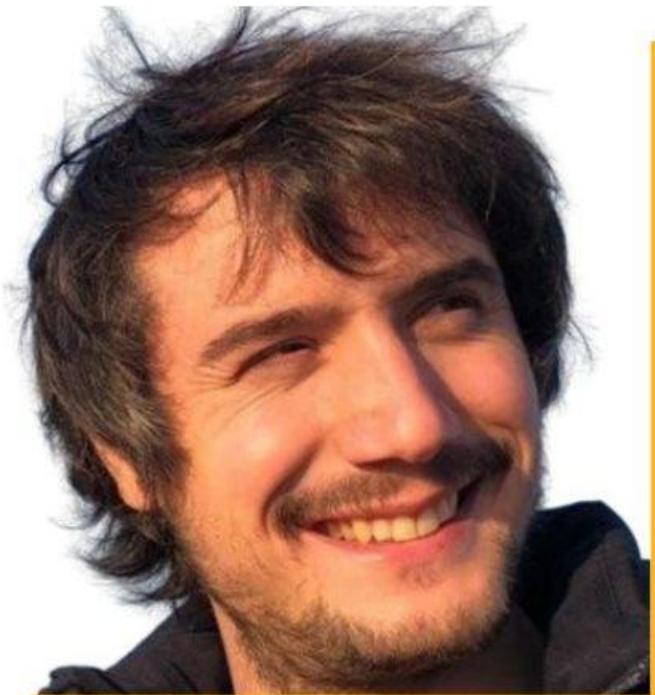


lonestone



[sfēir]

*“REX sur la création d'Auto PHP,
un agent autonome open source”*



Bonus

<https://sites.research.google/videopoet/>

https://magicvideov2.github.io/?utm_source=www.therundown.ai&utm_medium=newsletter&utm_campaign=the-gpt-store-is-here

https://dreamtuner-diffusion.github.io/?utm_source=www.therundown.ai&utm_medium=newsletter&utm_campaign=top-ai-researchers-earn-865k-year