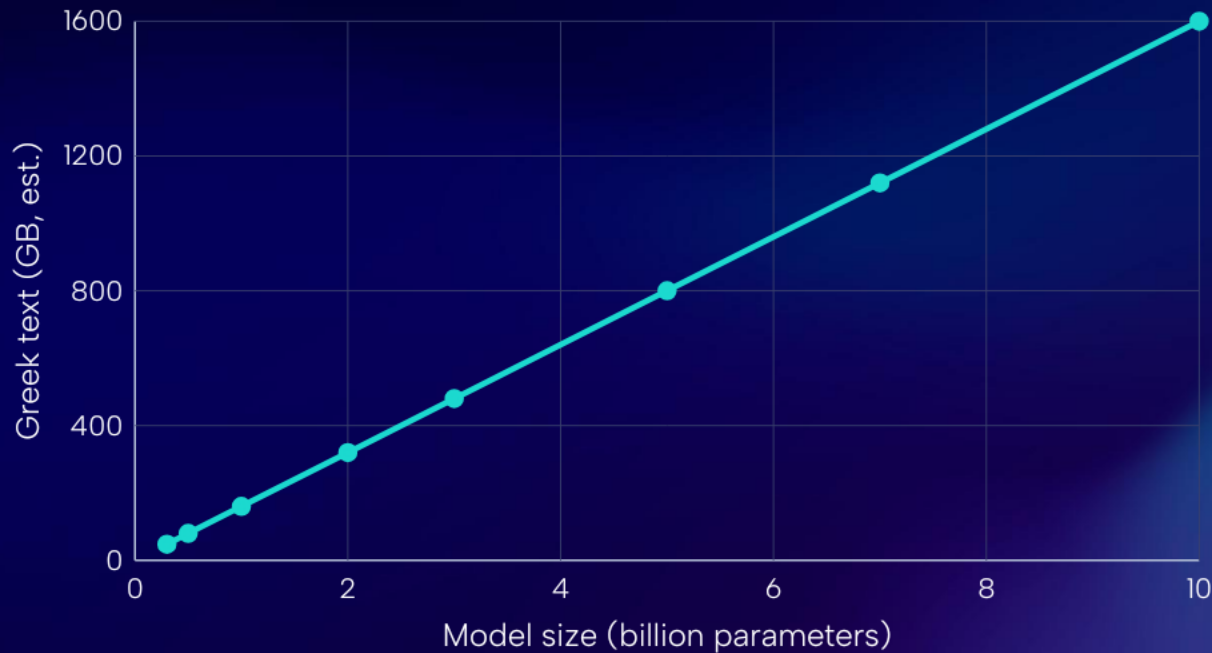


GlossAPI <γlo'sapi>

Open datasets as the foundation of AI

Foivos Karounos

Training data and model size

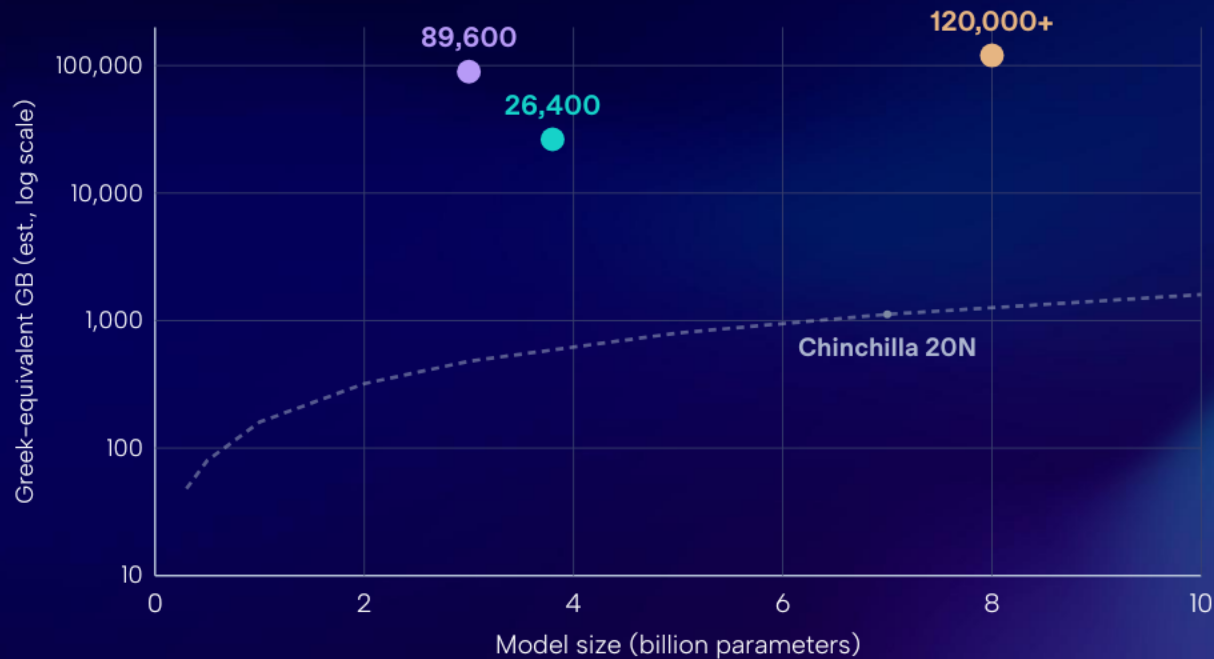


≈20

tokens per
parameter

7B parameters
≈1120 GB

Modern training scale



Phi-3-mini 3.8B

3.3T training tokens

SmoLM3 3B

11.2T training tokens

Llama 3 8B

15T+ training tokens

What share of the text is Greek?



eng
46%

rus
10%

deu
5%

fra
4%

Sino-Tibetan

zho
7%

por
2%

pol
2%

pes
1%

Modern Greek (1453-)

nld
1%

ell <
1%

```
labels=ell
text_chars=2.836049405e+11
parent=Indo-European
id=Indo-European/ell
```

spa 5%

ita
201

swe

ukr	0.488%
-----	--------

id=Indo-
lik
259%
0.18

Japanese

Austronesian

ind
1%

Austro-Asiatic

11000

Uralic

Afro-Asiatic

Figure 10

ron
1%

0.356%
nob
0.356%

[illegible]

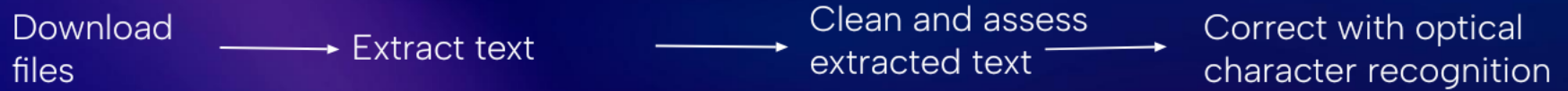
for

Uralic

Afro-Asiatic

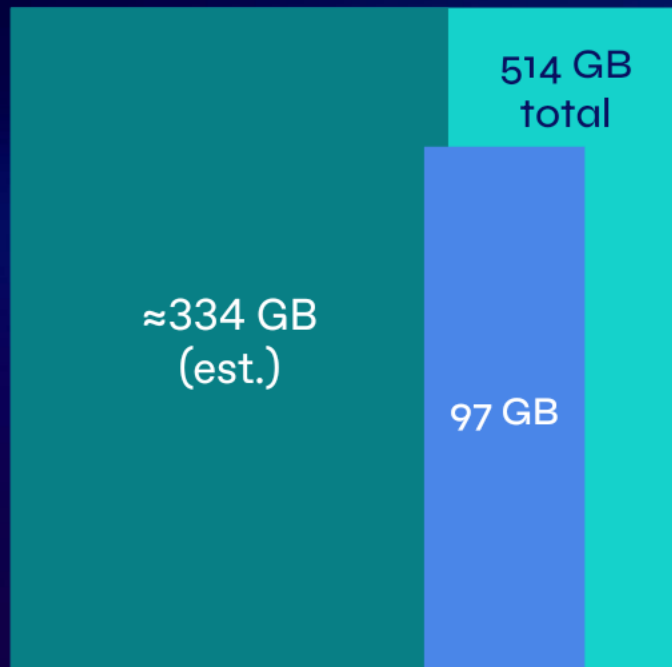
Figure 10

GlossAPI workflow



Large Greek-language datasets

-  HPLT 3.0
-  Quality ≥ 8
HPLT bin
-  GlossAPI



Greek Apertus

v0.5 / 8B

Swiss AI Initiative grant

Score (%)	Greek Apertus v0.5	Apertus Instruct	Krikri 1.0	Krikri 1.5
GreekMMLU	70.52	66.50	67.94	67.69
Greek IFEval	65.80	54.90	63.77	54.90
Greek GSM8K	59.67	56.18	71.72	73.01

Contact

glossapi.gr

glossapi.team@eellak.gr

