European Language Resource Coordination
Connecting Europe Facility

European Commission

The value of data for the development of top quality LT

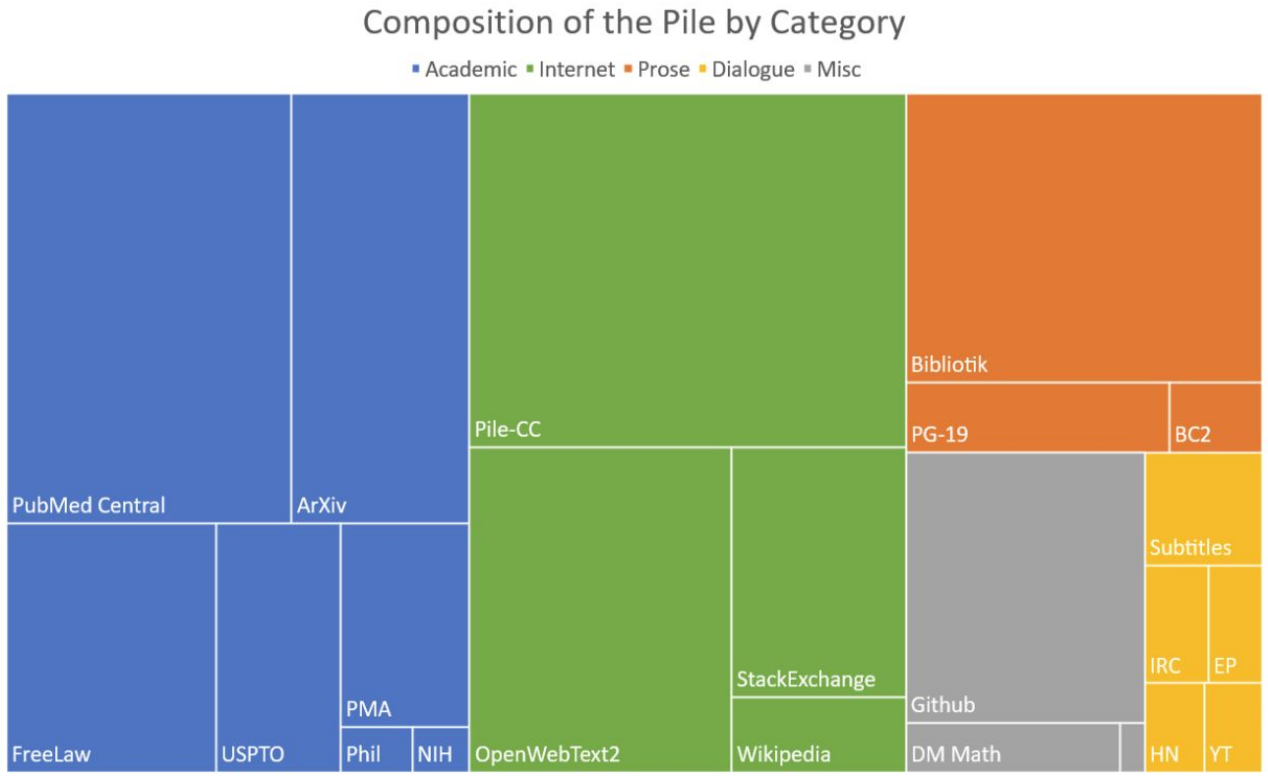
[Adam Feldmann University of Pécs]

Why we need more data for language models?

- Language models are the engine of language technology
- Era of large language models
- S.O.T.A solutions need more data
  (aka. Pains and promises of BERT,GPT-3, etc.)

How Much Data is Enough Data?

GPT-3, Wu Dao 2.0, Megatron-LM 530B



(Source: [2101.00027] The Pile: An 800GB Dataset of Diverse Text for Language Modeling (arxiv.org)

How Much Data is Enough Data?

- Hard to tell….but it is calculable now based on empirical researches.

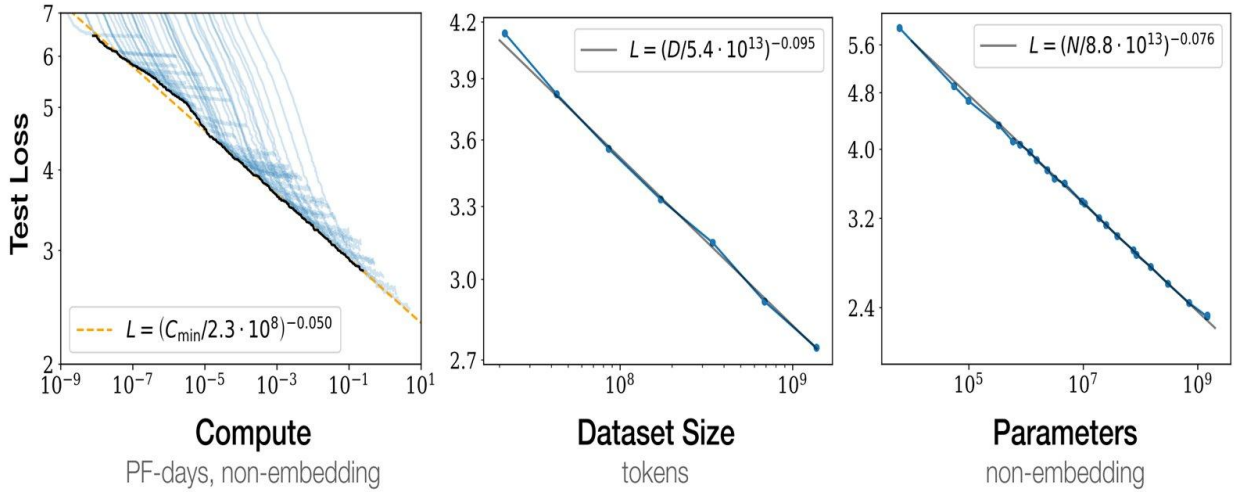
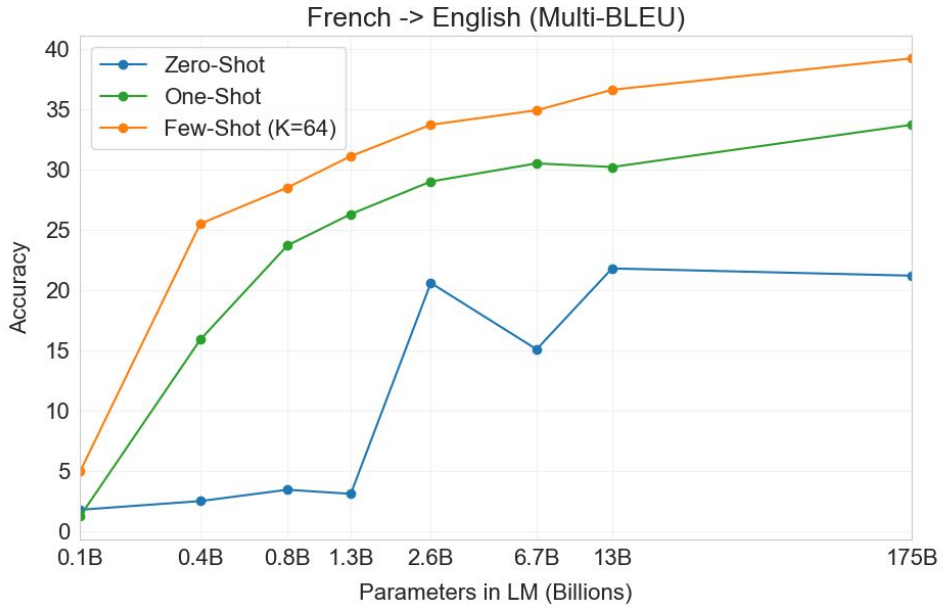


Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

(Source: 2001.08361.pdf (arxiv.org) Jared Kaplan et al; 2020; Scaling Laws For Neural Language Models

Machine Translation (MT) using Prompt Programming Paradigm



Few-shot

In addition to the task description, the model ses a few examples of the tasks. No gradient updates are performed.

1	*Translate English to French:*	*task description*
2	*sea oter => loutre de mer*	*examples*
3	*peppermint => menthe poivrée*	
4	*plush girafe => girafe peluche*	
5	*chese =>*	*prompt*

SCIFORCE

An unexpected beneficial phenomenon: Prompt programming

- New paradigm in AI.
- A new method for using multitask metalearner models like GPT.
- Originated from zero-shot or few-shots learning.
- Prompting allows the user to generate a very specific behavior from the system.

(Source:[2005.14165] Language Models are Few-Shot Learners (arxiv.org))

European Language Resource Coordination
Connecting Europe Facility

European Commission

Thank you for your attention!

Website: www.lr-coordination.eu
Twitter: @LR_Coordination
Facebook: www.facebook.com/EuropeanLanguageResourceCoordination
LinkedIn: www.linkedin.com/in/lrcoordination
Email: info@lr-coordination.eu