Data Scaling Laws in NMT: The Effect of Noise and Architecture

Yamini Bansal[1,2] Behrooz Ghorbani[2] Ankush Garg[2] Biao Zhang[3]
Maxim Krikun[2] Colin Cherry[2] Behnam Neyshabur[2] Orhan Firat[2]

[1]Harvard University

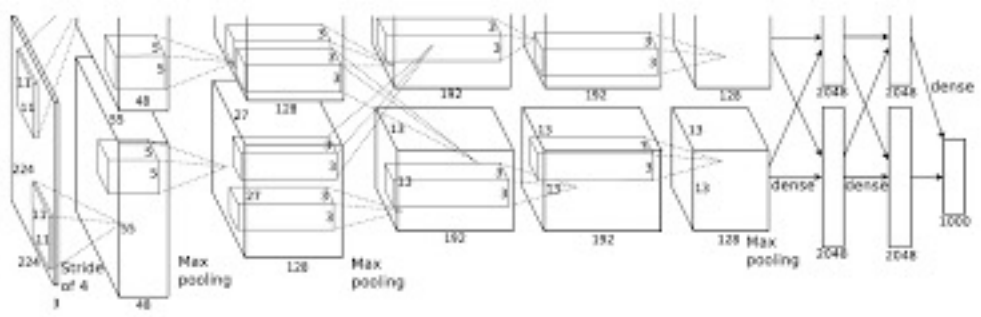
[2]Google

[3]University of Edinburgh

Harvard John A. Paulson
School of Engineering
and Applied Sciences

More data is better

Reliable way to improve performance: Add more data!

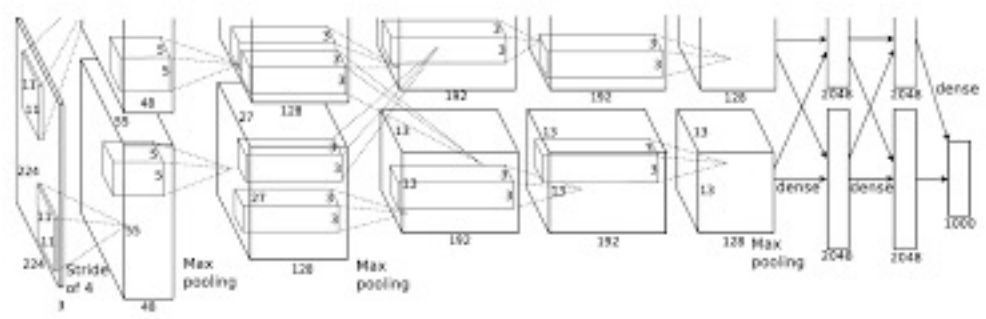


AlexNet

GPT3

More data is better

Reliable way to improve performance: Add more data!



AlexNet

GPT3

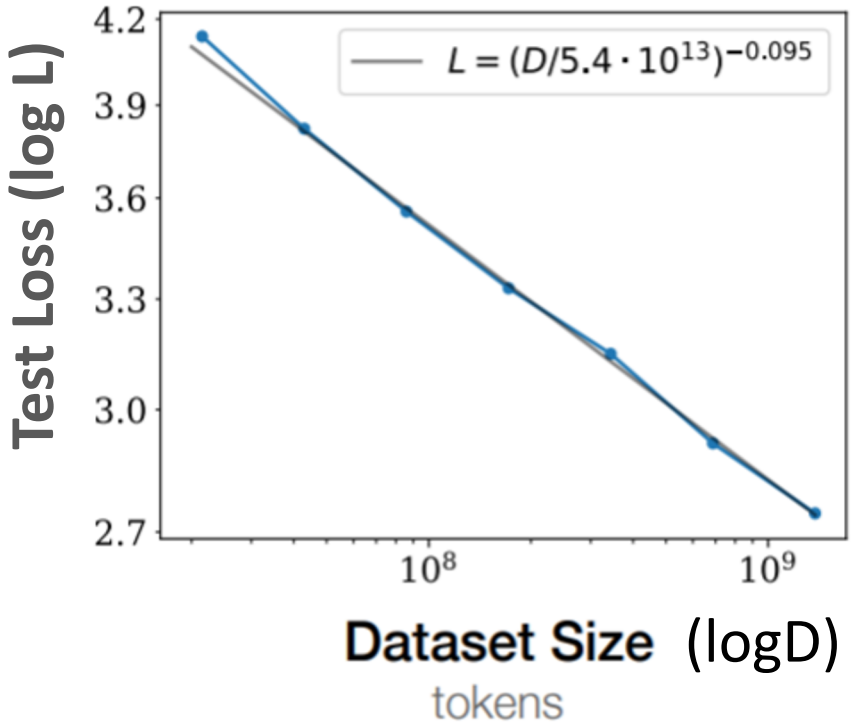
Sample-efficiency: How does performance *scale* with increasing data?

Measuring sample-efficiency: Neural scaling laws

Measuring sample-efficiency: Neural scaling laws

Test Loss (L) empirically scales as a power law in D = Training Dataset size

$$L = \alpha \frac{1}{D^p}$$



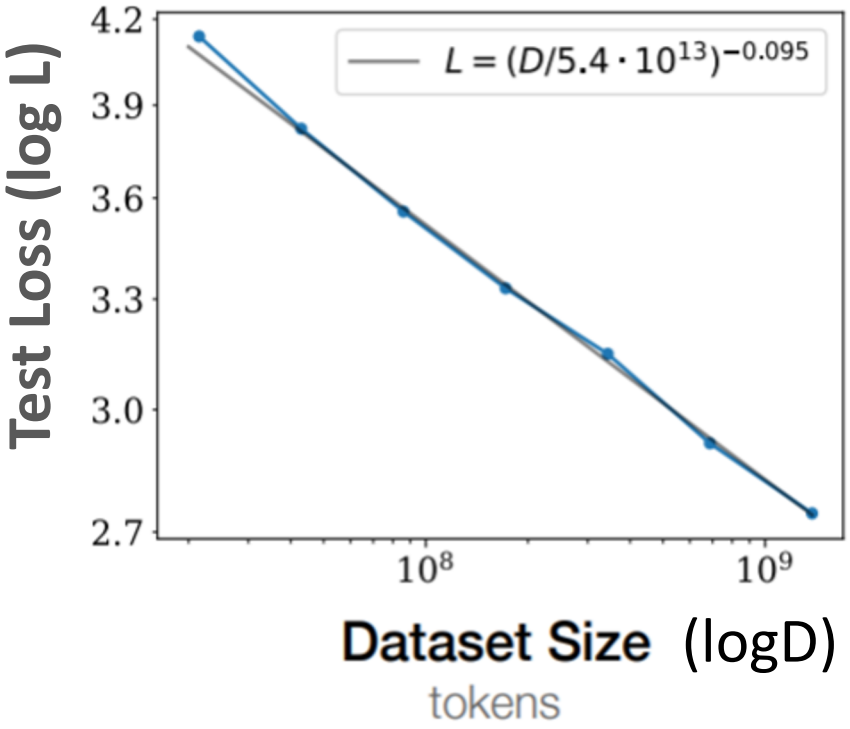
From Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J. and Amodei, D., 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

See also Hestness et. al. (2017), Rosenfeld et. al. (2019)

Measuring sample-efficiency: Neural scaling laws

Test Loss (L) empirically scales as a power law in D = Training Dataset size

$$L = \alpha \frac{1}{D^p}$$



Exponent p summarizes the 'sample-efficiency'

From Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J. and Amodei, D., 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

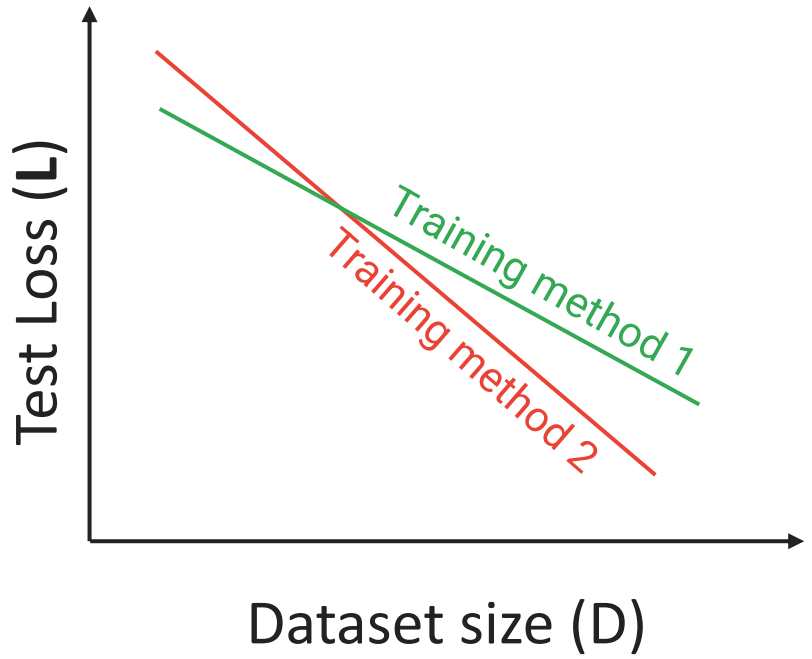
See also Hestness et. al. (2017), Rosenfeld et. al. (2019)

Why study data scaling laws?

Why study data scaling laws?

Practical Reasons

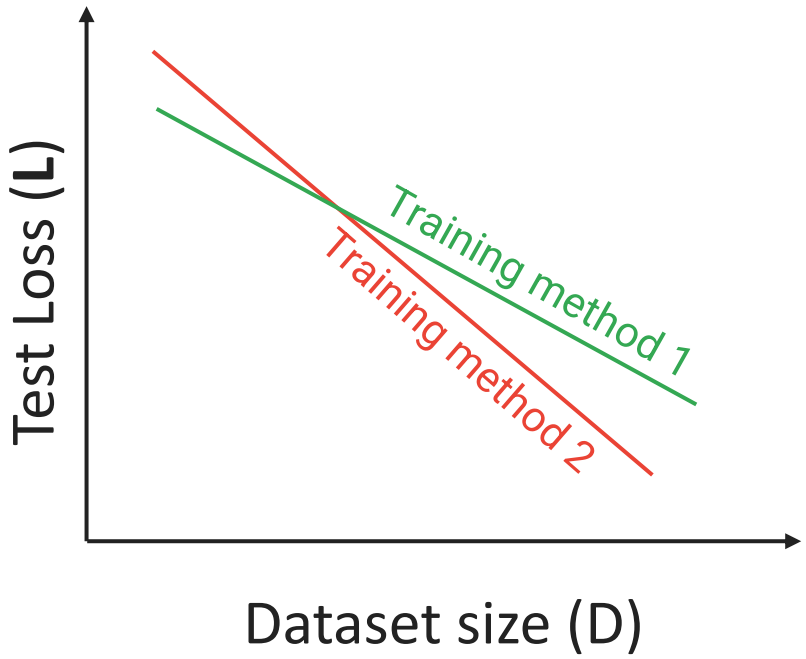
- Make predictions for larger scale experiments
- Comparisons at single point are not enough



Why study data scaling laws?

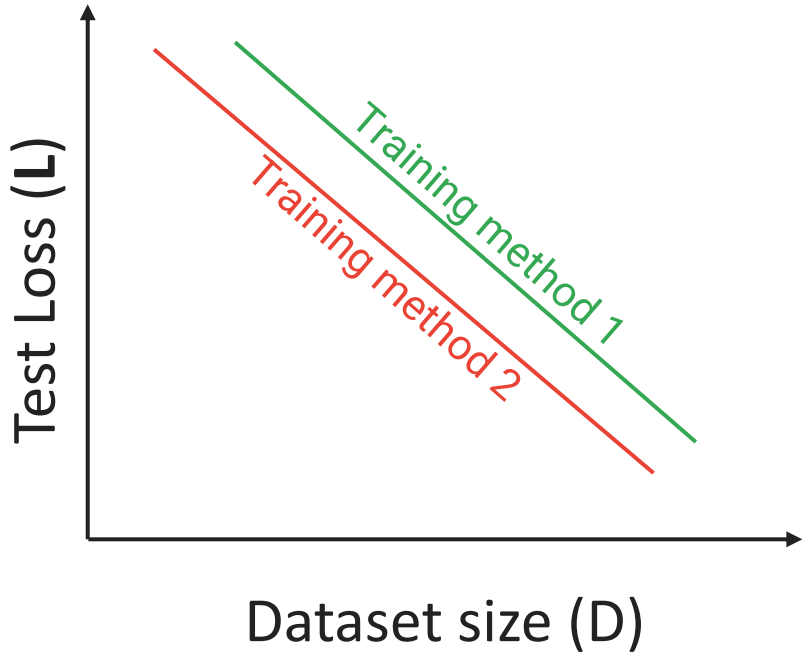
Practical Reasons

- Make predictions for larger scale experiments
- Comparisons at single point are not enough



"Theoretical" Reasons

- If many methods scale similarly, can we understand why?
- How can we do better?



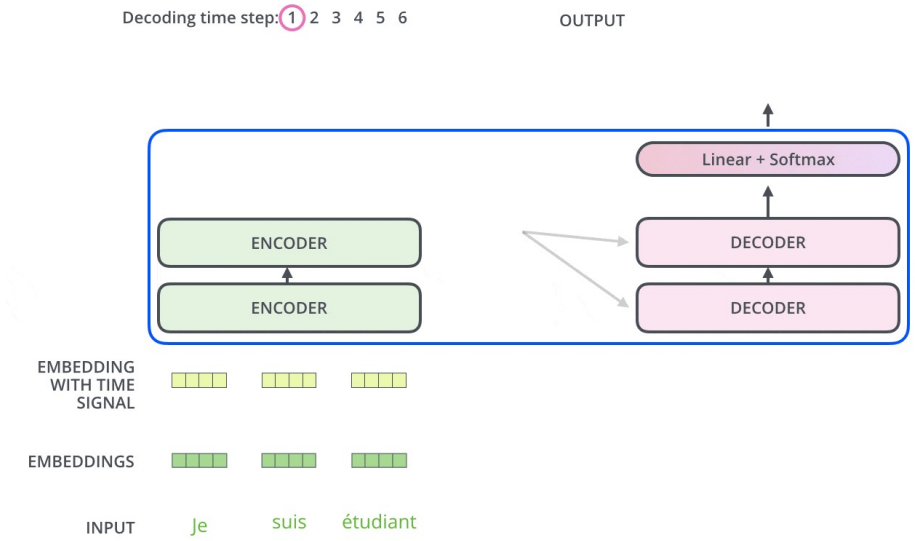
Our contributions

Which aspects of the training setup affect the data scaling empirically?

Our contributions

Which aspects of the training setup affect the data scaling empirically?

- Scaling laws for Neural Machine Translation (NMT)
  - Encoder-Decoder Transformers
  - English → German



Translation

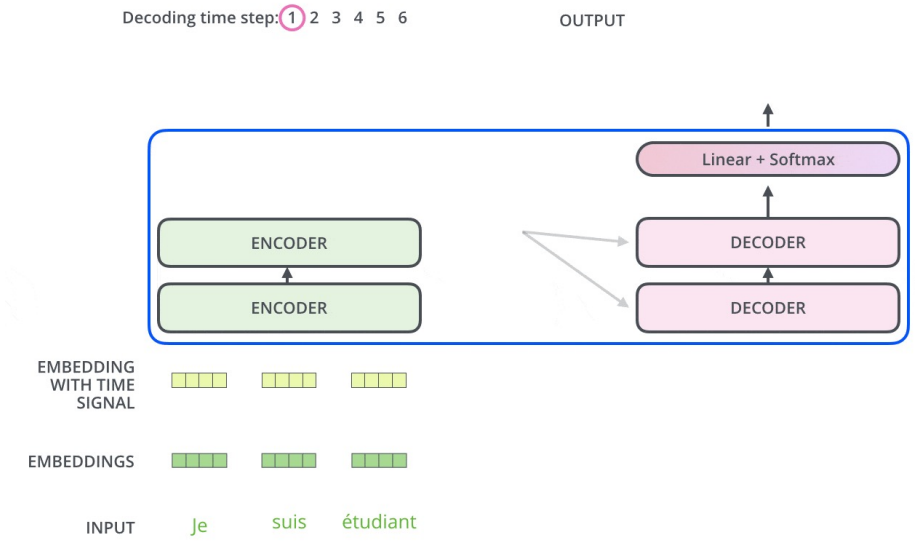
Source: https://jalammar.github.io/illustrated-transformer/

Our contributions

Which aspects of the training setup affect the data scaling empirically?

- Scaling laws for Neural Machine Translation (NMT)
  - Encoder-Decoder Transformers
  - English → German
- Interventions to training setup
  - Change architecture
  - Change noise in training distribution

Important practical tools



Translation

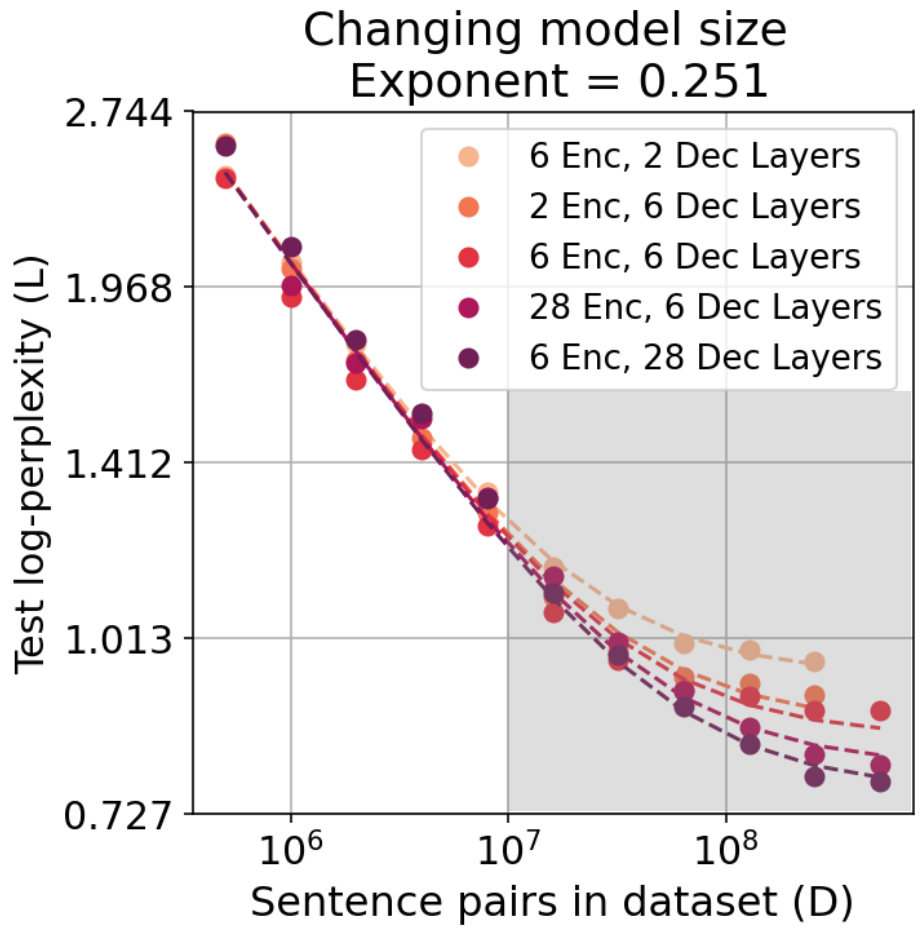
Source: https://jalammar.github.io/illustrated-transformer/

NMT Data Scaling Law

NMT Data Scaling Law

We fit the scaling law

$$\text{Loss} = \alpha(1/D + C_m)^p$$



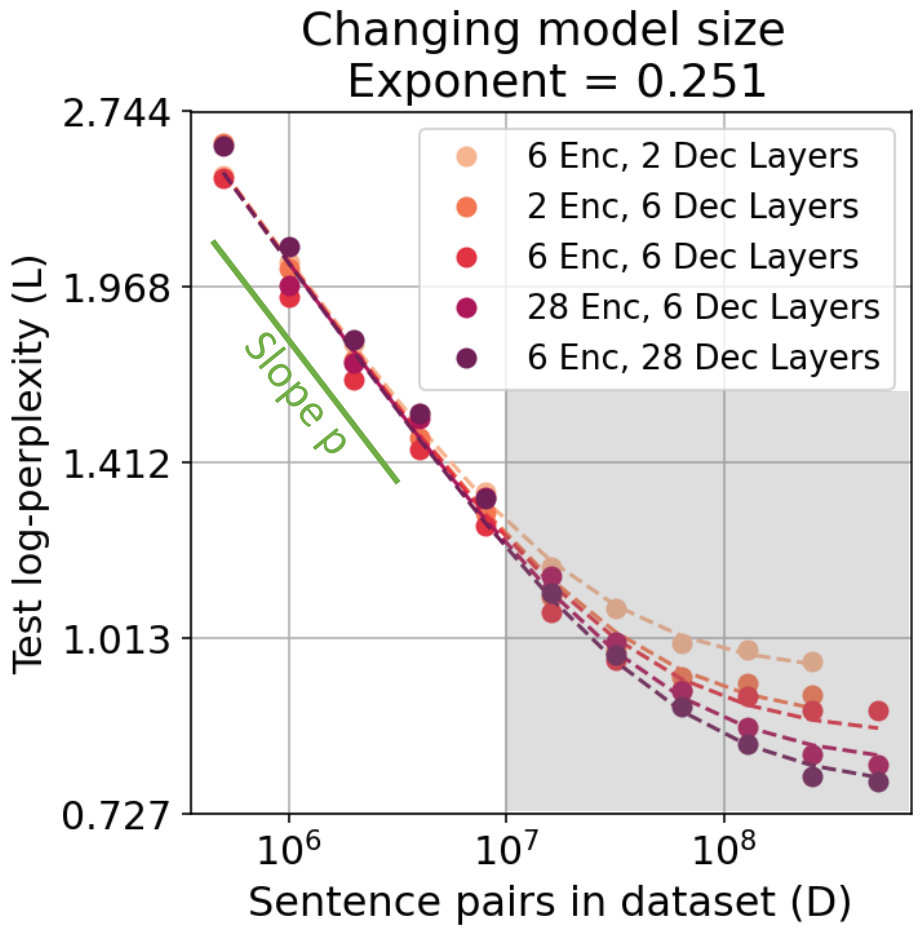
NMT Data Scaling Law

We fit the scaling law

Data Limited Regime

$$\text{Loss} = \alpha(1/D + C_m)^p$$

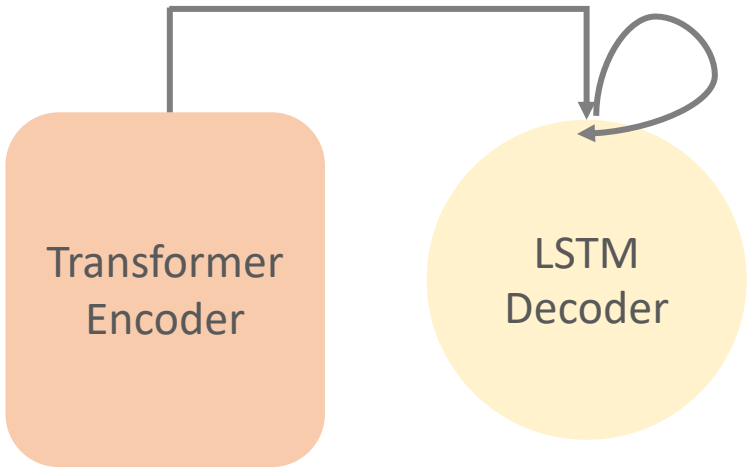
- Increasing model size doesn't help
- Exponent independent of encoder-decoder depth ratio



Effect of Architecture

Effect of Architecture

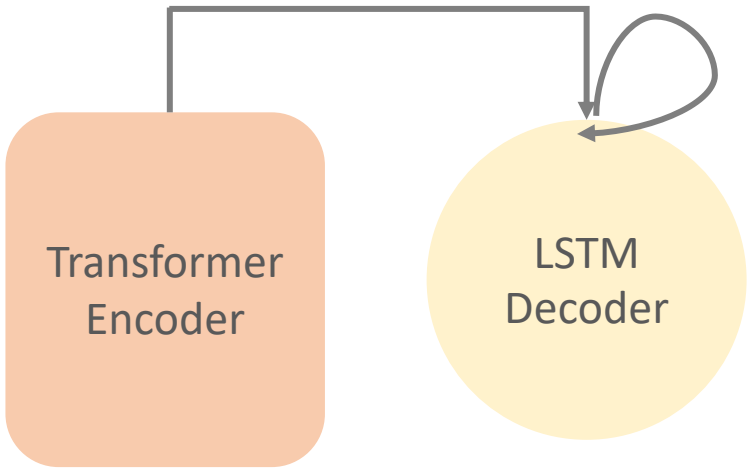
Transformer Encoder – LSTM Decoder



Common in industry applications

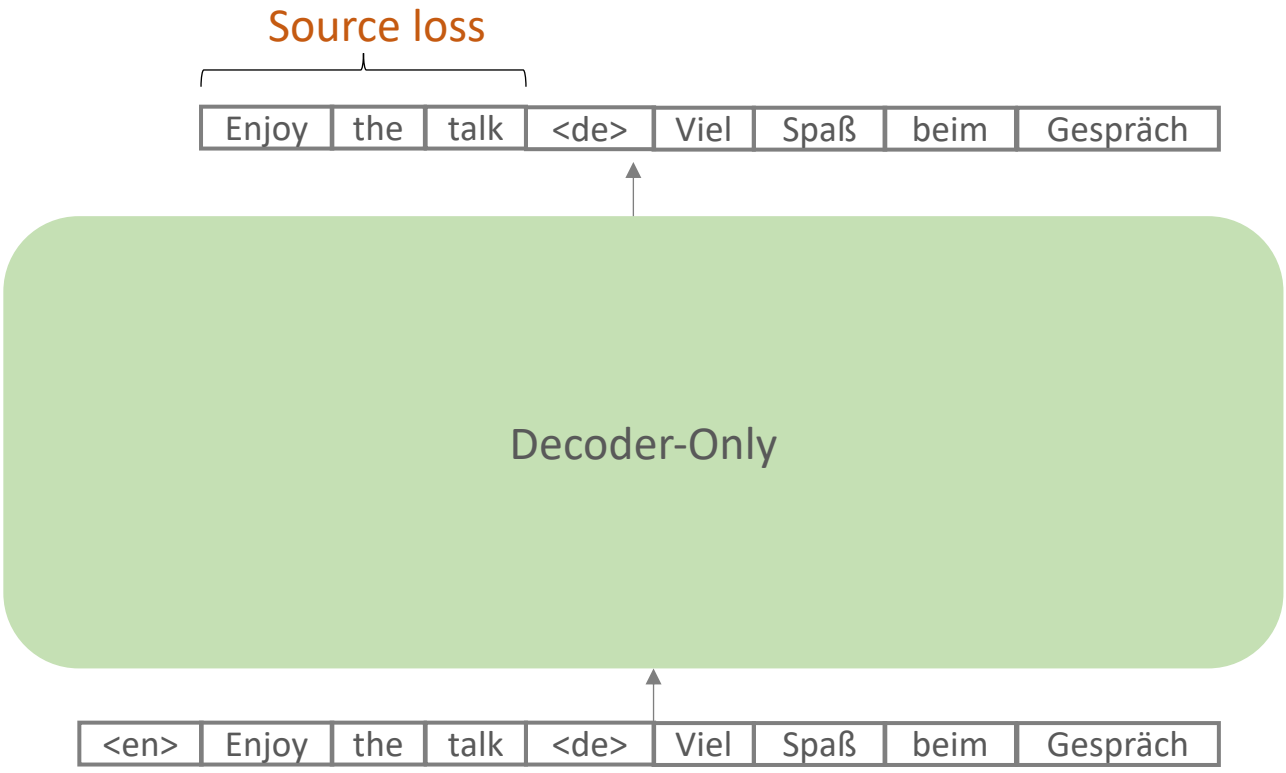
Effect of Architecture

Transformer Encoder – LSTM Decoder



Common in industry applications

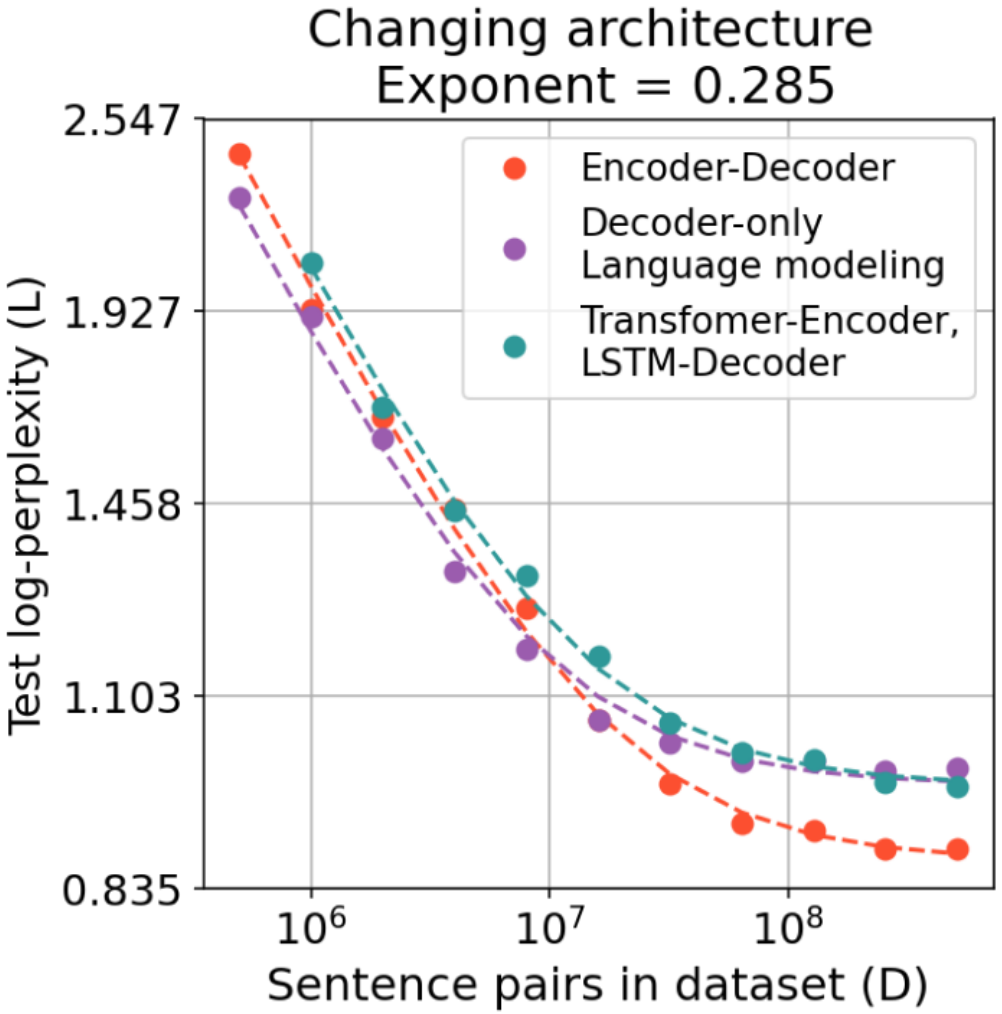
Decoder-only Transformer (GPT)



Same setup as GPT models

Effect of Architecture

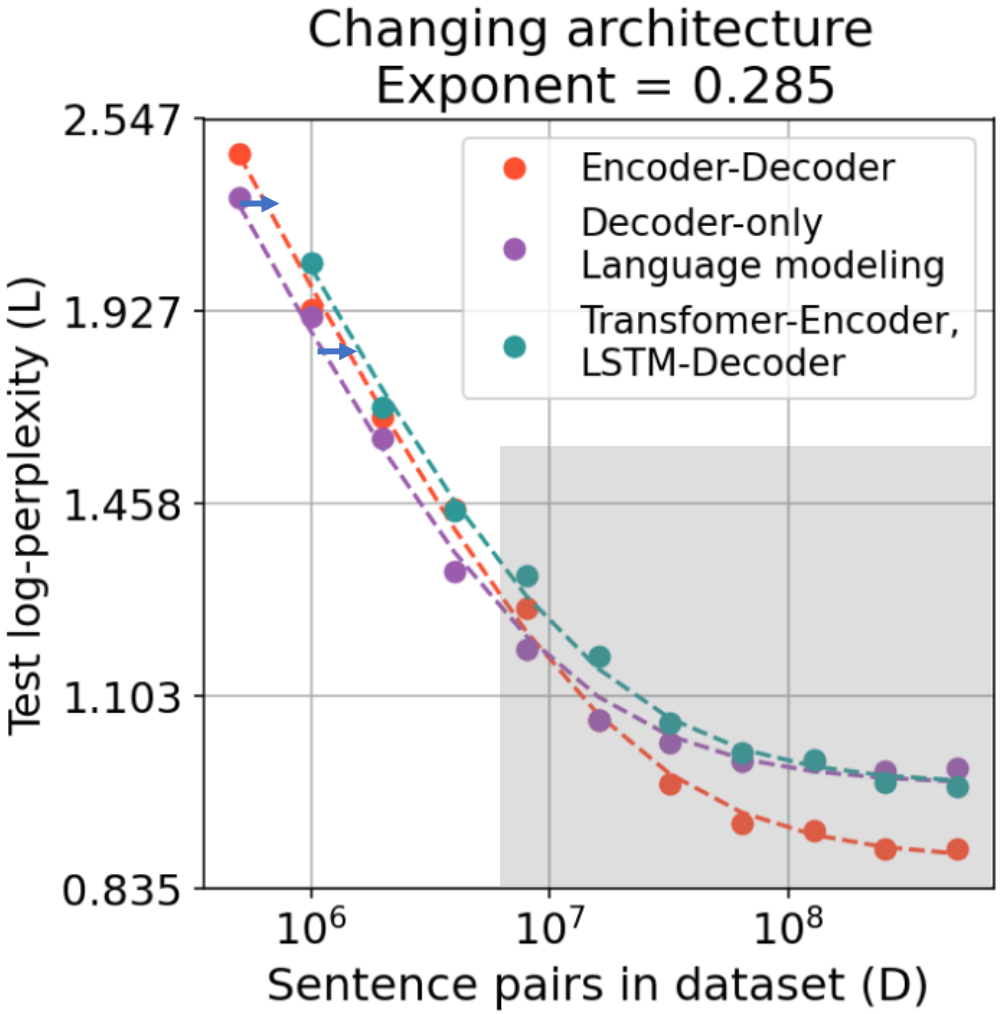
Loss = $\alpha_m(1/D + C_m)^p$



Effect of Architecture

$$\text{Loss} = \alpha_m(1/D + C_m)^p$$

- Common exponent p
- We can compensate for a weaker architecture by adding more data

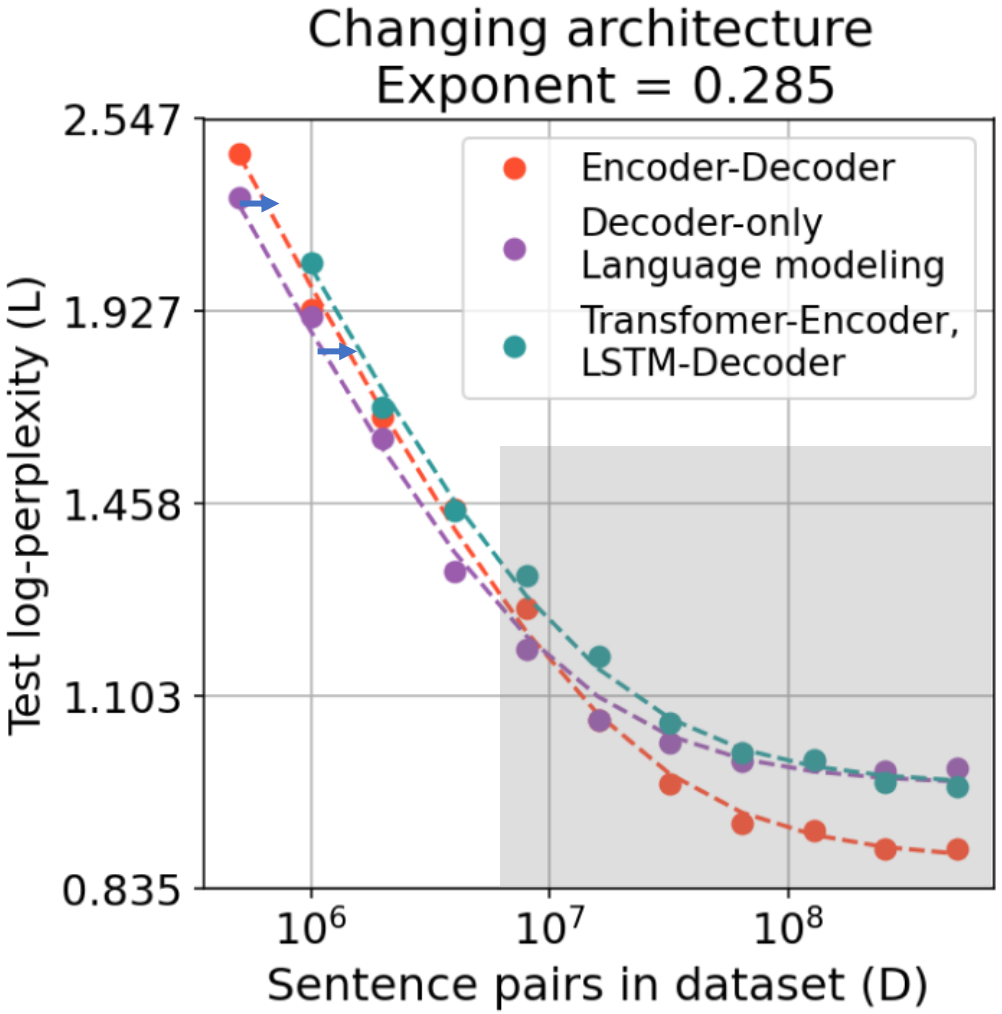


Effect of Architecture

$$\text{Loss} = \alpha_m(1/D + C_m)^p$$

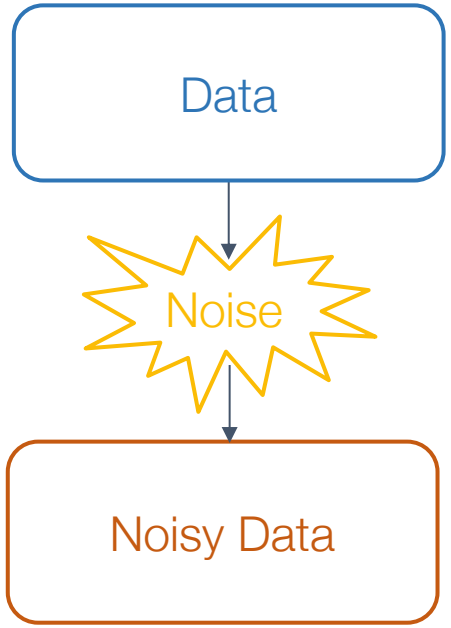
- Common exponent p
- We can compensate for a weaker architecture by adding more data

If you have other priorities (eg: compressibility), you can choose a worse architecture, by training it with more data



Effect of Noise

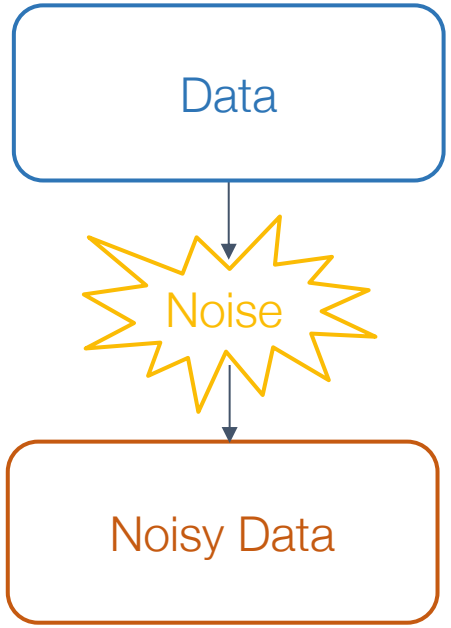
Adding synthetic noise



Source Noise: <2de> Me%t all our products Seiten, die auf „Siemens (Einheit)“ verlinken - brand-feuer.de
Target Noise: <2de> These allergy-sufferers often wonder if purifiers are good for airborne allergies. Profus=on ist4eine Kühr/noe Desig~er Muskel Zelle

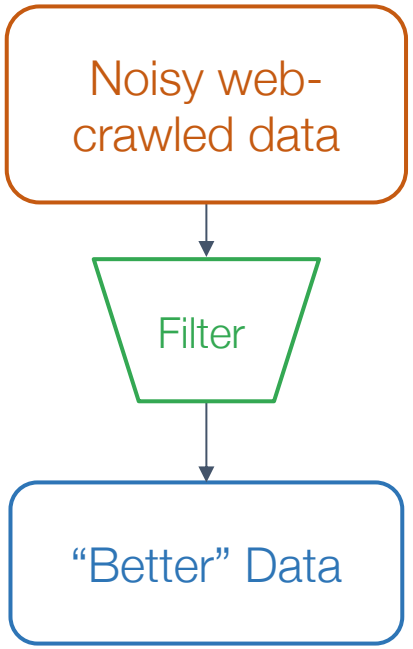
Effect of Noise

Adding synthetic noise



Source Noise: <2de> Me%t all our products Seiten, die auf „Siemens (Einheit)“ verlinken - brand-feuer.de
Target Noise: <2de> These allergy-sufferers often wonder if purifiers are good for airborne allergies. Profus=on ist4eine Kühr/noe Desig~er Muskel Zelle

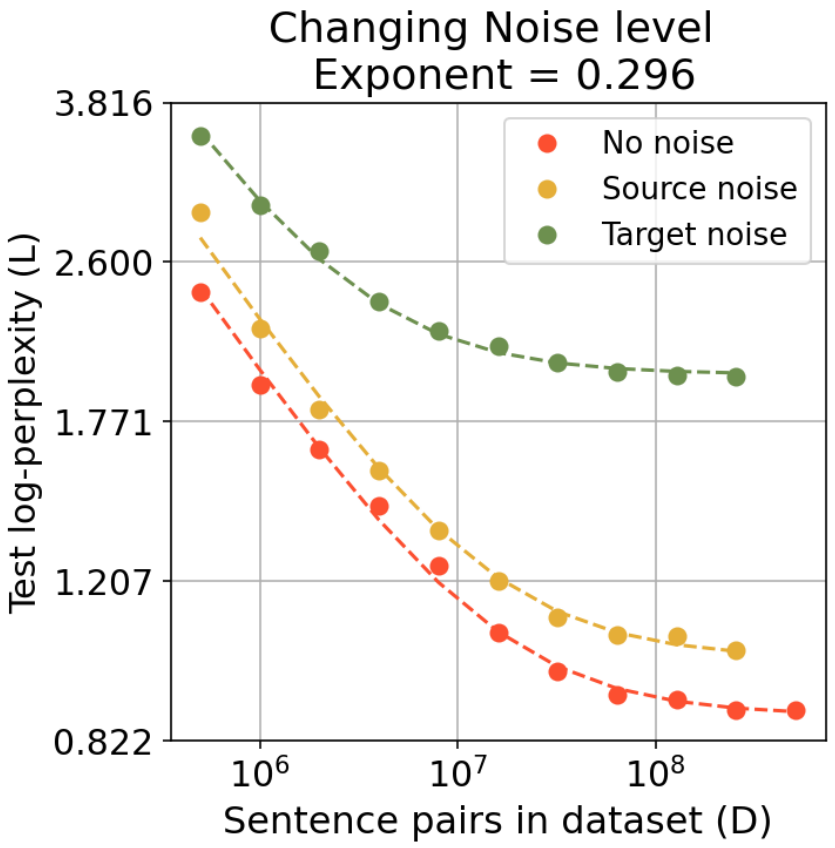
Filtering – Subtracting nosie



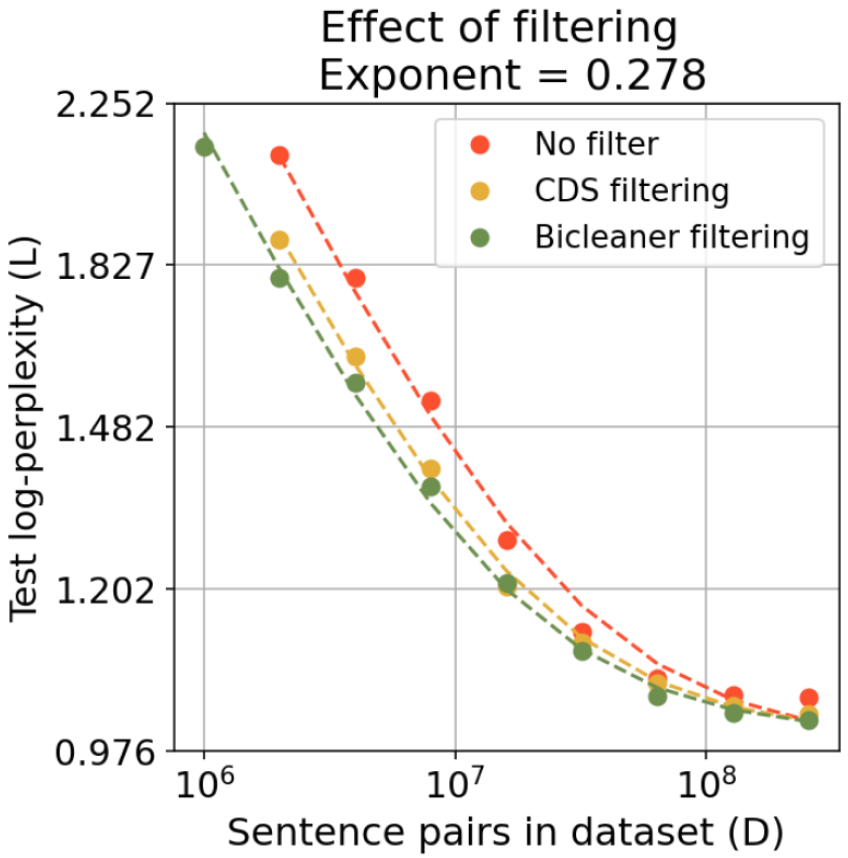
- Wrong language
- Language model score for fluency
- Unaligned sentences
- Too many special characters

Effect of Noise

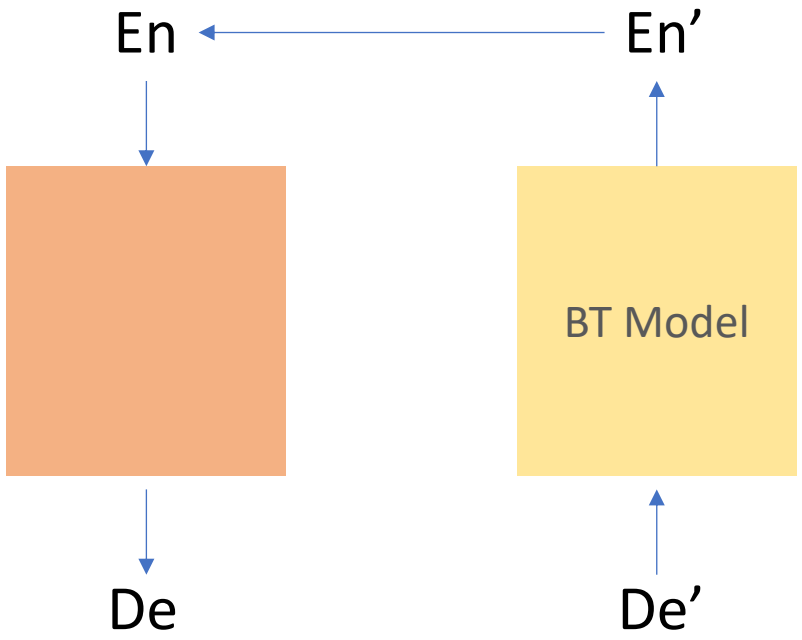
Adding synthetic noise

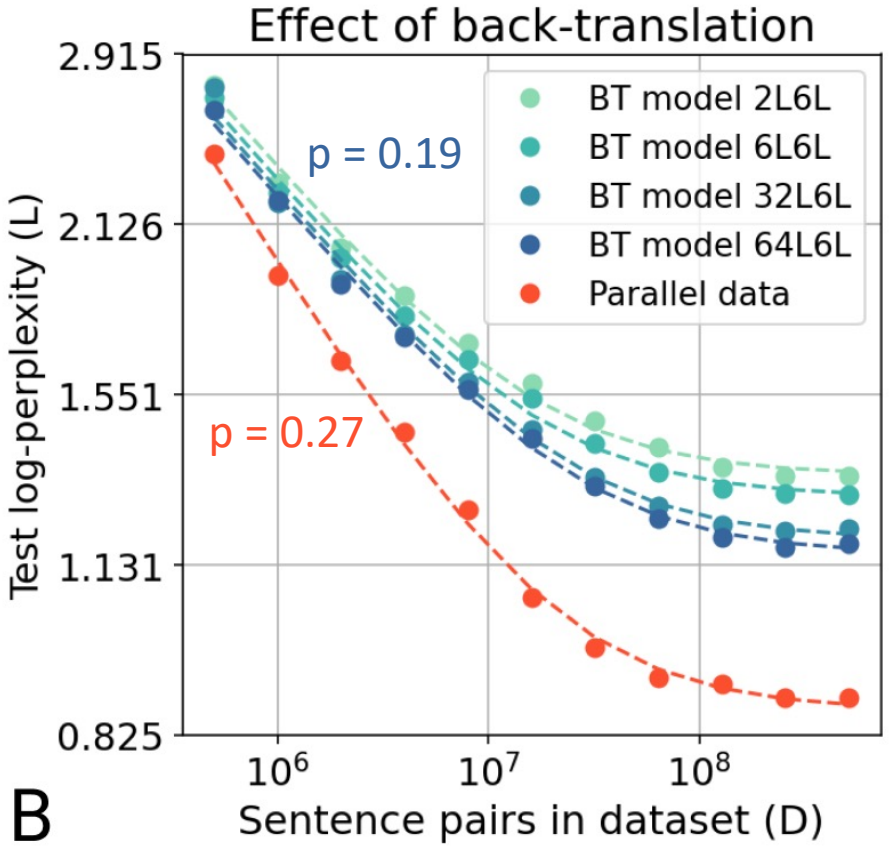


Filtering



Change in exponent: Back-translation





Takeaways and open questions

Practical:

- Scaling laws - rigorous tool to drive practical trade-offs
- You can compensate for certain “worse” choices like noise and sub-optimal architecture by adding a constant fraction of more data

Theoretical:

What is the “inductive bias” that keeps exponent similar?