Proč umí velké jazykové modely česky, slovensky, německy…, i když je to pořádně nikdo neučí?

Jindřich Libovický

13. 12. 2023

Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

unless otherwise stated

Outline

1. How do language models work
2. Selection of language model types
3. NLP Tasks solved with language models
4. Generative models
5. From LMs to assistants
6. Stochastic parrots and other problems
7. Research at ÚFAL MFF CUNI

How do LMs work

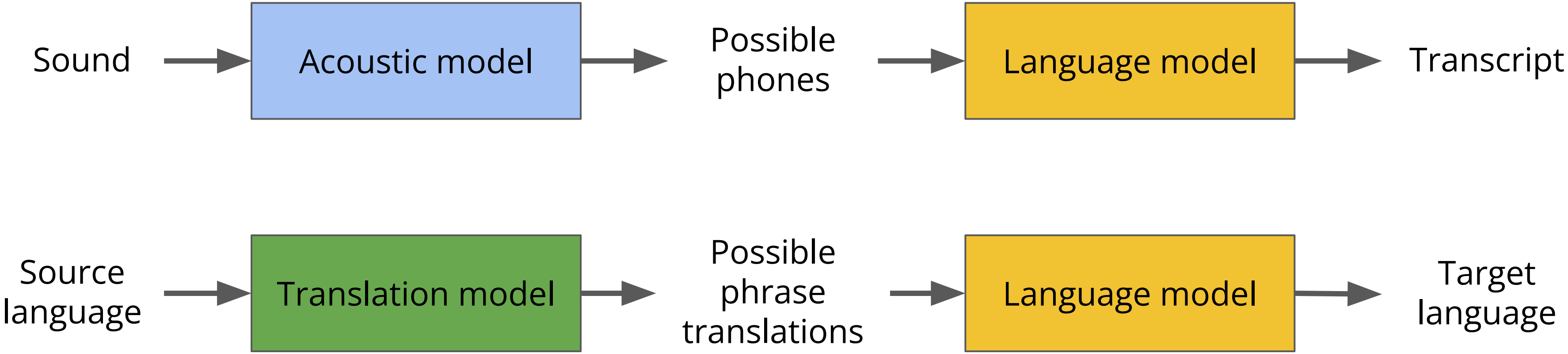
Language Model

Estimate the probability of a word / sentence / text in a context.

Where do LMs come from

Since 1990’s

An important component in speech recognition and machine translation



Sound → Acoustic model → Possible phones → Language model → Transcript

Source language → Translation model → Possible phrase translations → Language model → Target language

LMs used as...

1. Models of what is good/bad in a language
2. Representation learning models
3. Generative models

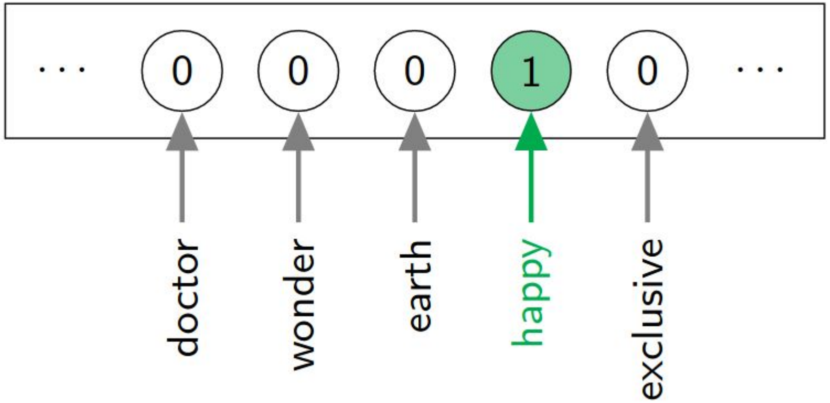
Types of Language Models

Neural Language Models

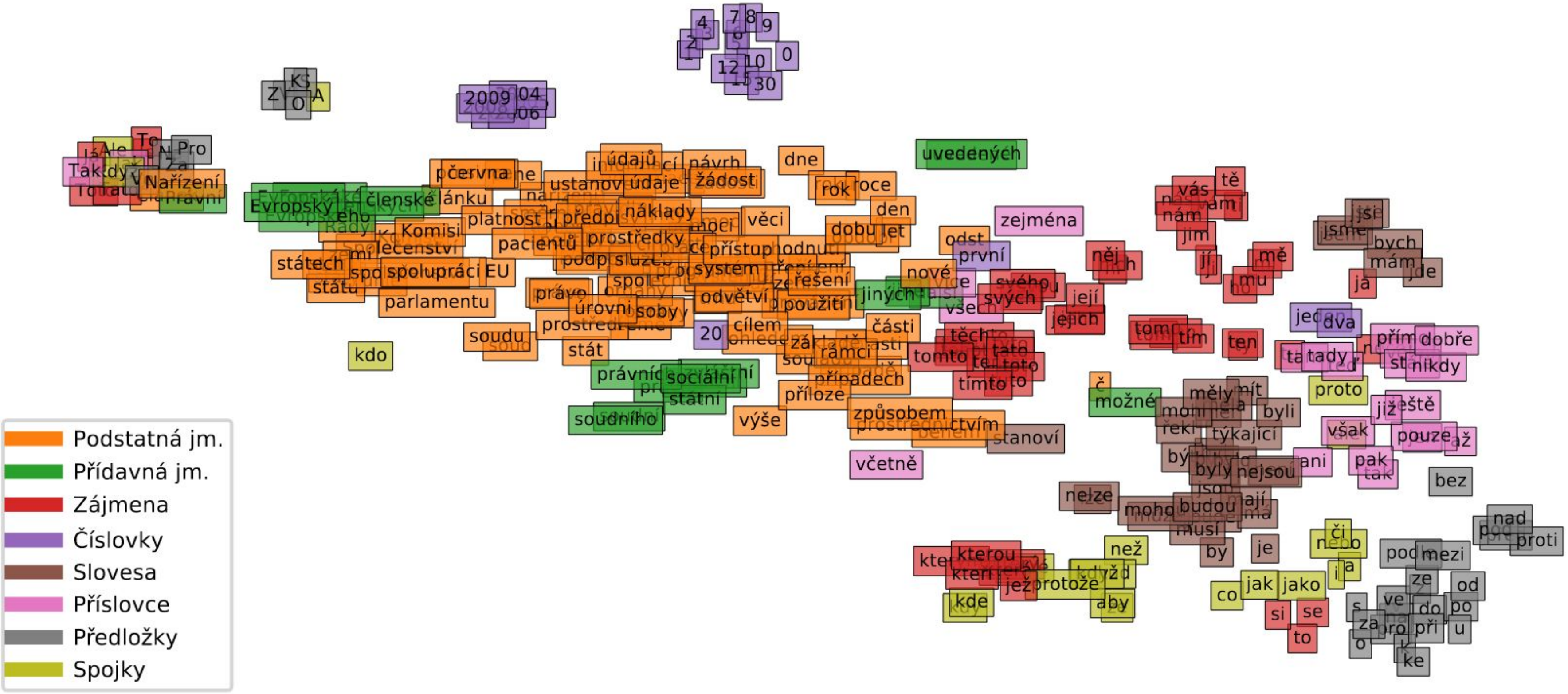
1. **Encoder-decoder modely**
    - Machine translation, text summarization
2. **Encoder-only modely**
    - BERT, RoBERTa, ALBERT, ...
    - Pretrained representation for downstream tasks
3. **Decoder-only = generative models**
    - GPT, ChatGPT

Text Processing with Neural Networks

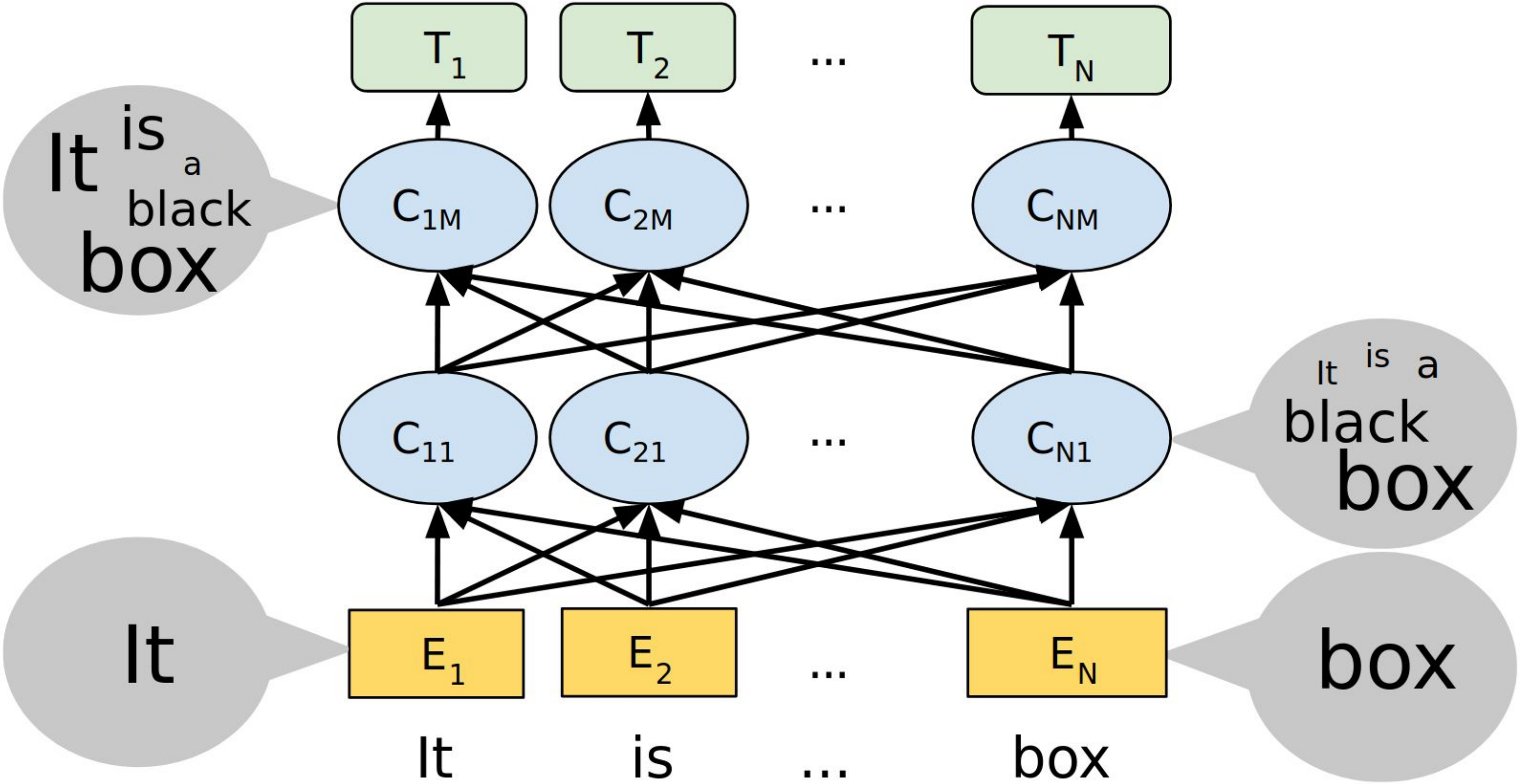
- NNs work with real numbers, **text is discrete**
- Words segmented to **tokens** (subwords)
- Tokens represented by vectors in continuous space (embeddings)
  - parameters of the NN - trainable
- Output is normalized and interpreted as **probability distribution** over token vocabulary



Visualization of Embeddings from an MT System

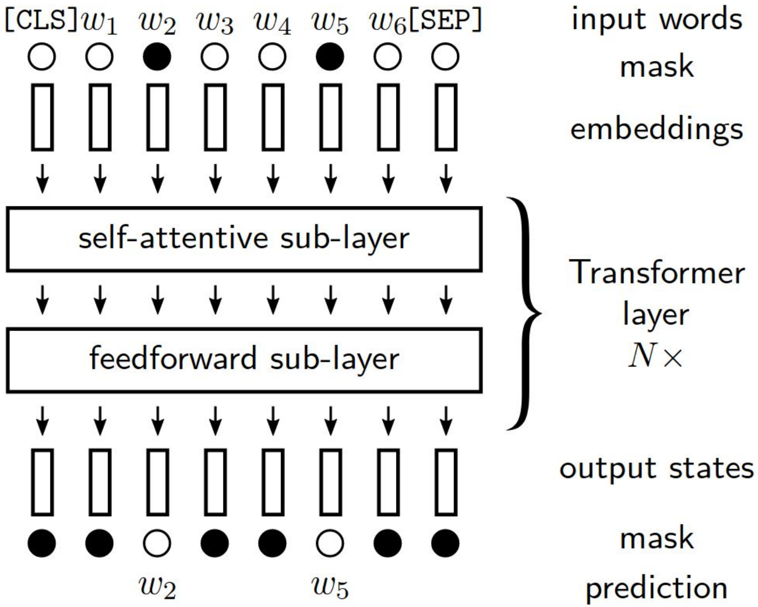


Intermediate Representations



Encoder-only: BERT

- The original Transformer for MT
  Encoder + Decoder
- For **representations**,
  encoder is enough
- Training without decoder
  Masked Language Modeling



Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171-4186).

Multilingual Pretraining

- The same as monolingual, but **many languages at once**
- Benefits from language similarity for **low-resourced** languages
- Multilingual tasks, e.g., training data **filtering** for machine **translation**

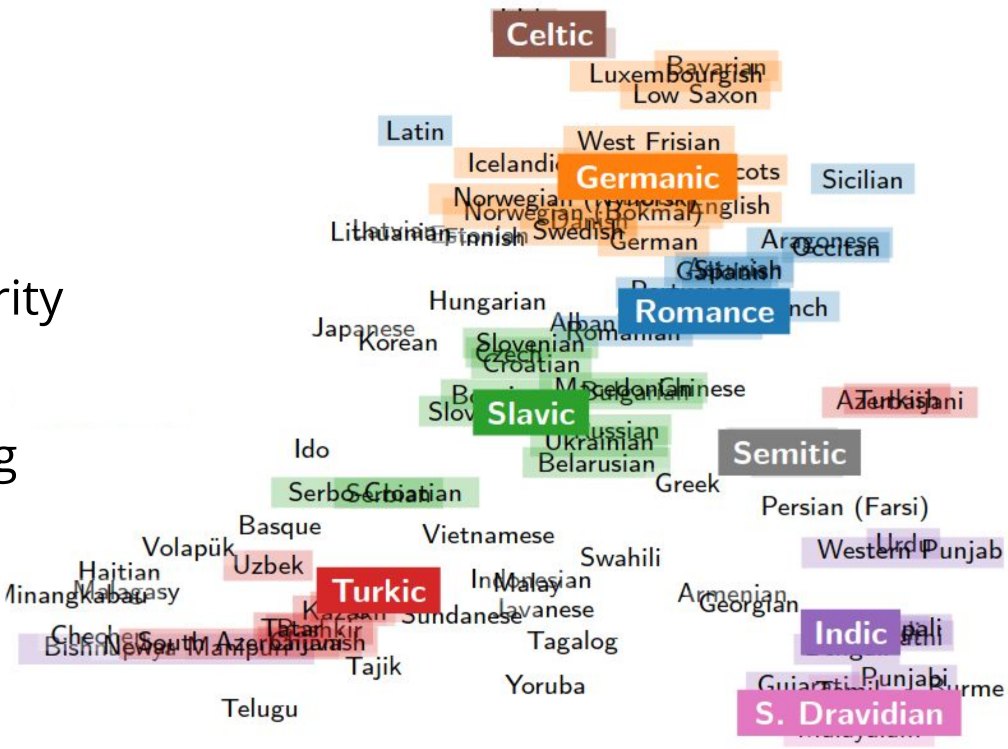


Image: Libovický, J., Rosa, R., & Fraser, A. (2020, November). On the Language Neutrality of Pre-trained Multilingual Representations. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 1663-1674).

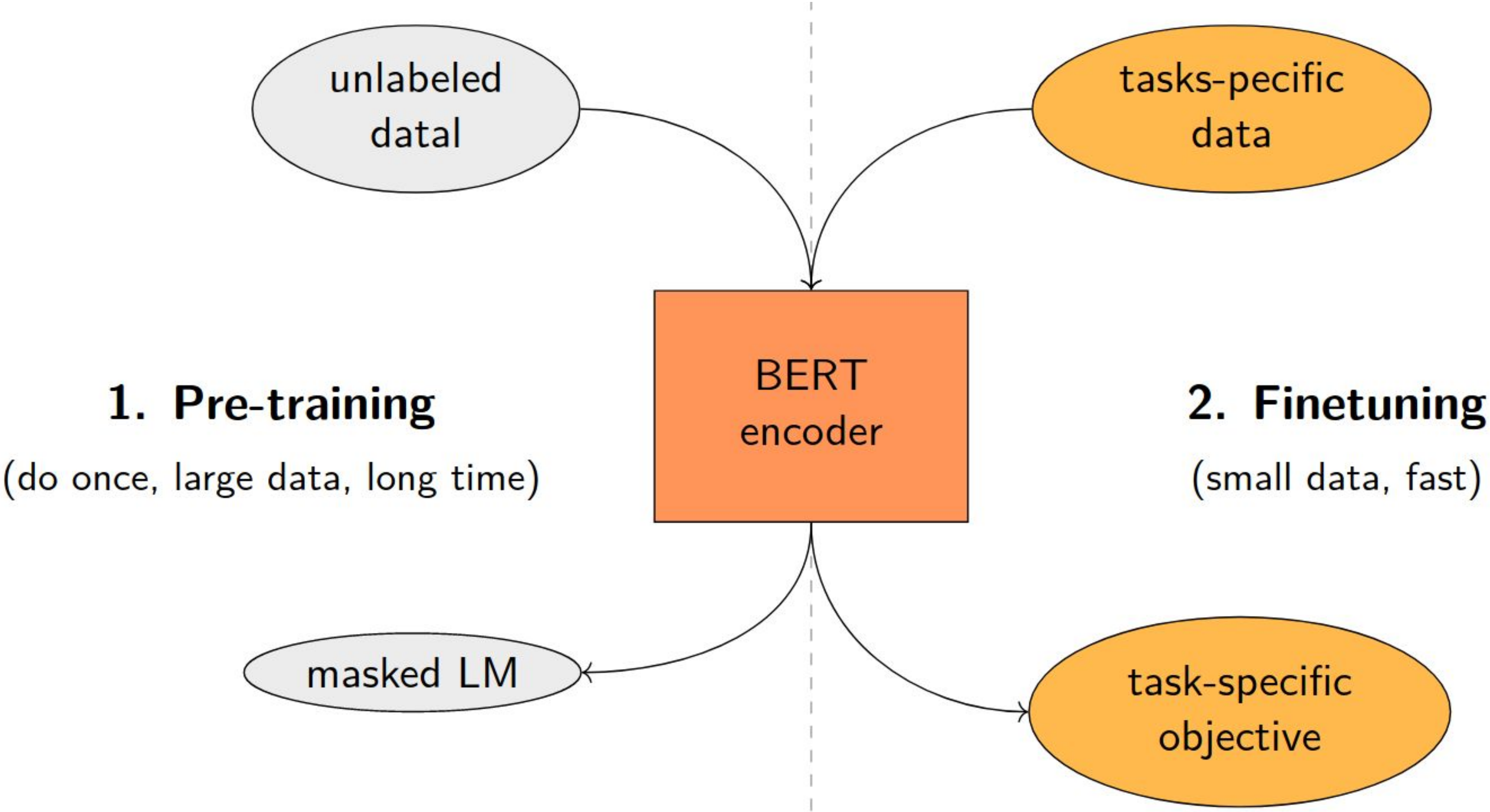
Notable BERTs

- BERT by Google, 2018
  110M parameters, 16GB of text
- RoBERTa by Facebook AI, 2019
  123M 160 GB of text
- XLM-R by Facebook AI, 2019
  125M parameters, 2.5TB of text
- RobeCzech by ÚFAL, 2020
  125M parameters, 80GB of text

Parameters for base setup,
Large setup twice as many params

NLP Tasks Solved using LMs

Pretrain and Finetune Paradigm



unlabeled datal

tasks-pecific data

1. Pre-training

(do once, large data, long time)

BERT encoder

2. Finetuning

(small data, fast)

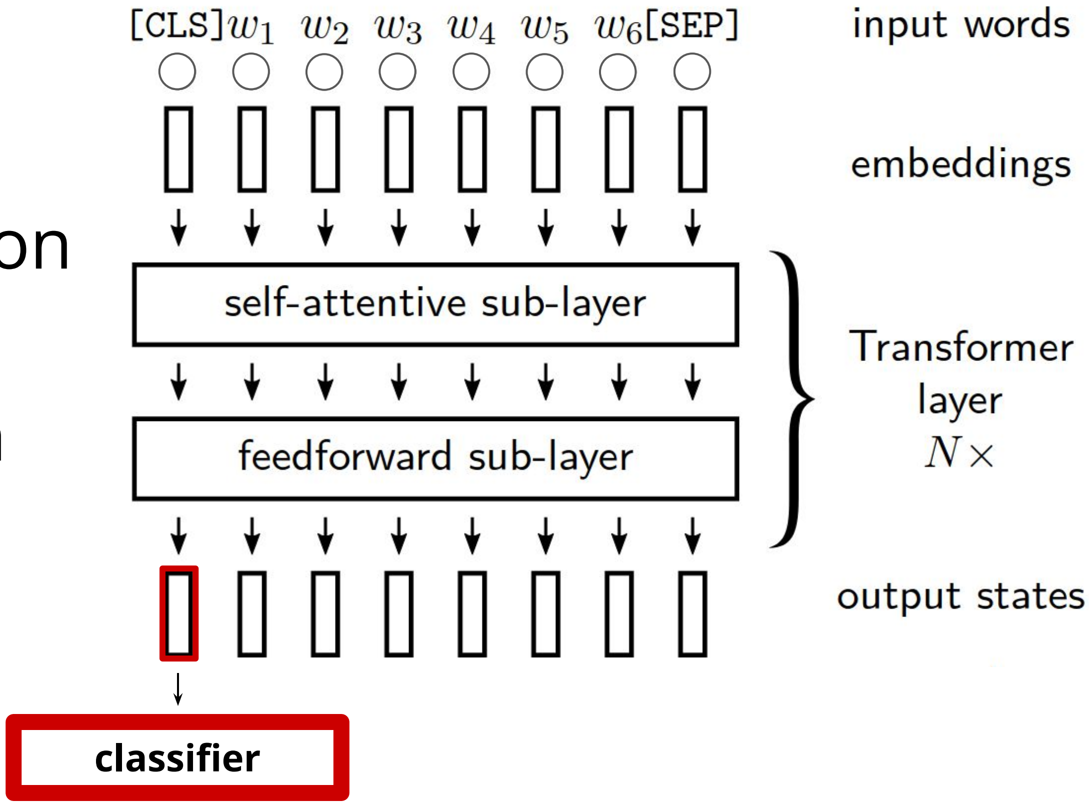
masked LM

task-specific objective

Classification

- Sentiment analysis
- Hate speech detection
- Spam detection
- Plagiarism detection

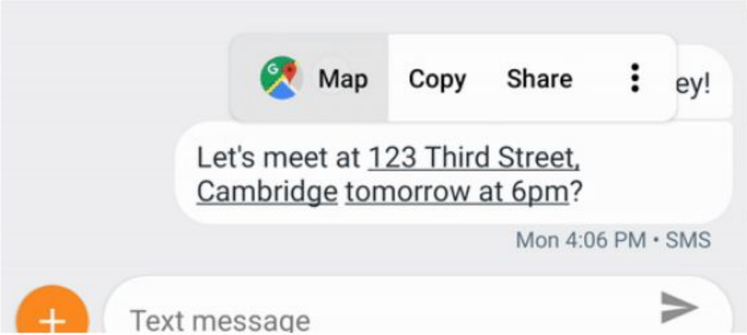
…



Named Entity Recognition (1)

- Part of information extraction pipeline
  - Entity linking (e.g., matching Wikipedia articles)
  - Coreference resolution

    Whom does pronoun "they" refer to?
    Who is "the president" in a text?
- Indexing text for search
- Direct use in smart devices



Let's meet at 123 Third Street, Cambridge tomorrow at 6pm?

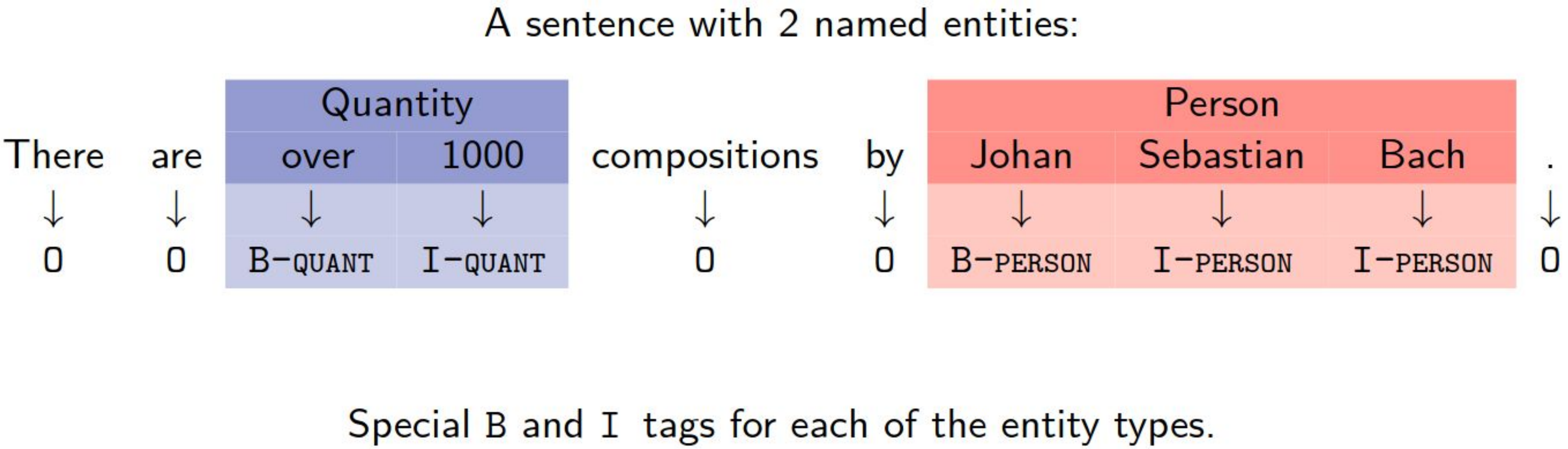
Mon 4:06 PM • SMS

Text message

NER used to create links in text to different apps.

Image source: Google AI Blog. `https://ai.googleblog.com/2018/08/the-machine-learning-behind-android.html`

Named Entity Recognition (2)

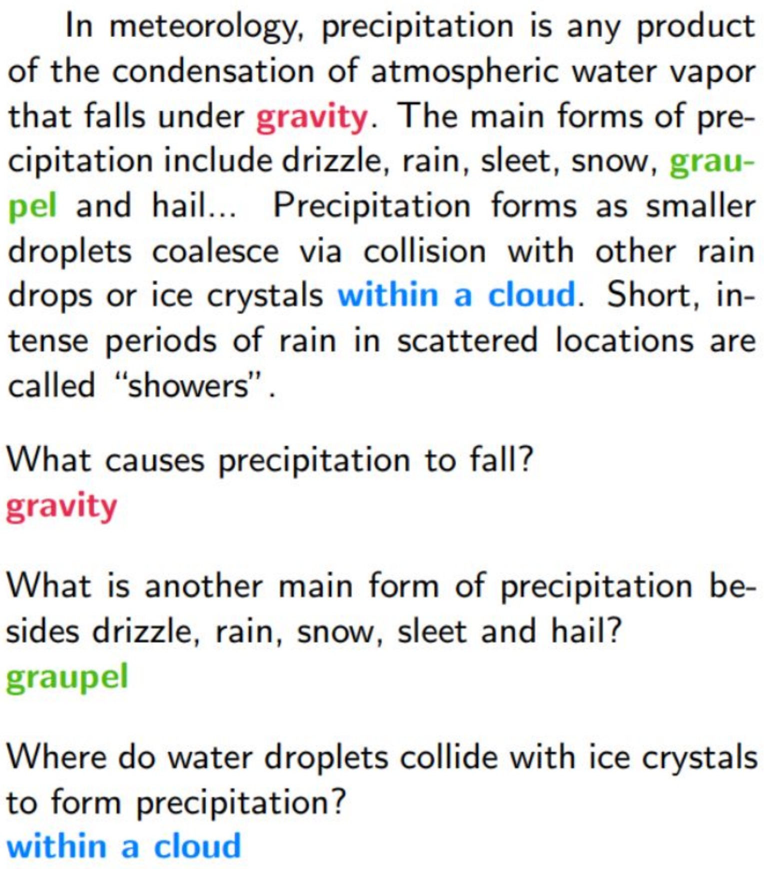


A sentence with 2 named entities:

		Quantity				Person			
There	are	over	1000	compositions	by	Johan	Sebastian	Bach	.
↓	↓	↓	↓	↓	↓	↓	↓	↓	↓
O	O	B-QUANT	I-QUANT	O	O	B-PERSON	I-PERSON	I-PERSON	O

Special B and I tags for each of the entity types.

Answer Span Selection (1)



In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under gravity. The main forms of precipitation include drizzle, rain, sleet, snow, graupel and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals within a cloud. Short, intense periods of rain in scattered locations are called "showers".

What causes precipitation to fall?
gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?
graupel

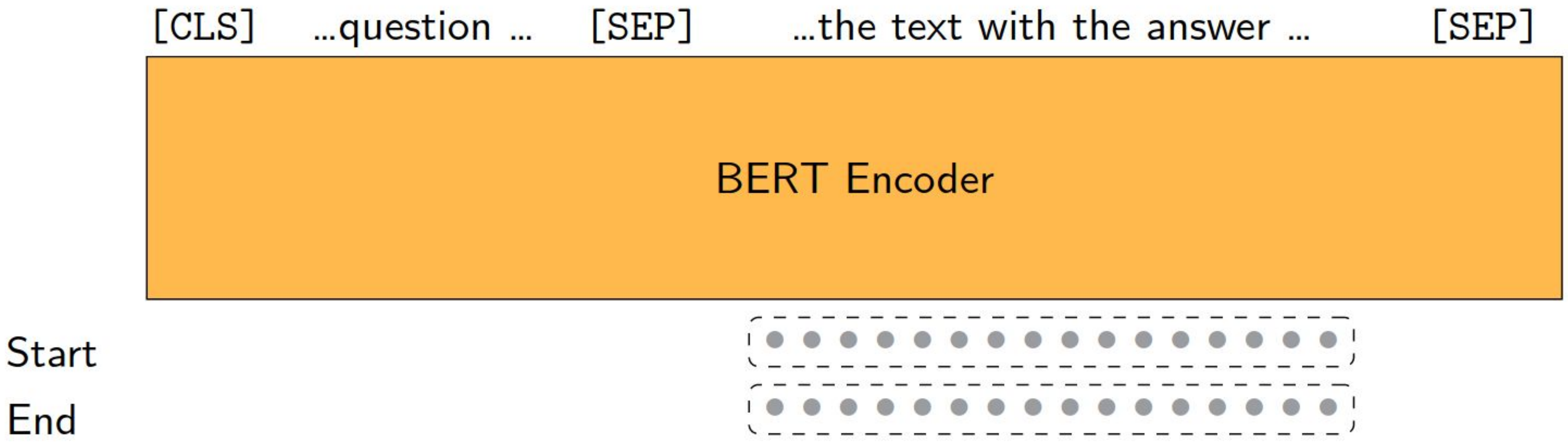
Where do water droplets collide with ice crystals to form precipitation?
within a cloud

- Input text with facts (e.g., a Wikipedia article)
- A factual question
- Model searches for an answer in the text

Rajpurkar, P., Zhang, J., Lopyrev, K., & Liang, P. (2016, November). SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing* (pp. 2383-2392).

Answer Span Selection (2)

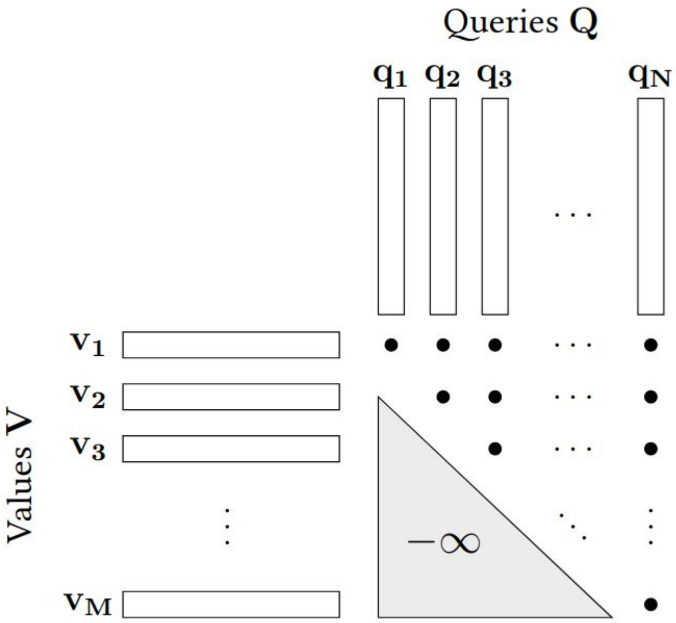
Just throw **everything into BERT**: both the text and the question.



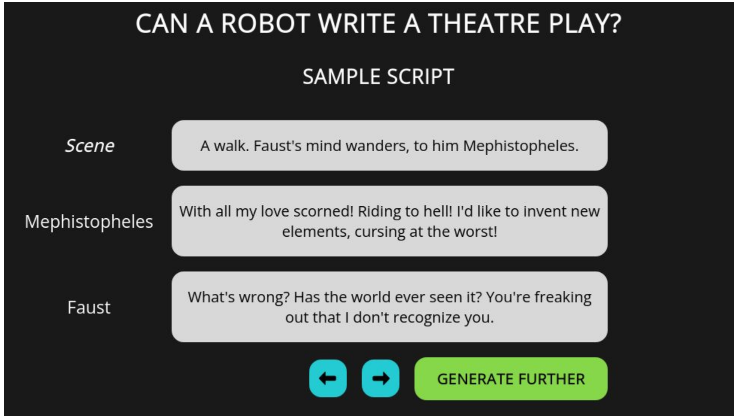
Generative Models

Decoder-only Models

- **Decoder** — just like encoder, but at training time masked **not to attend to the future**
- Training objective = **predict the next word** based on the previous words
    - Prompt provided from outside
    - Already generated text



Generating Any Text...



CAN A ROBOT WRITE A THEATRE PLAY?

SAMPLE SCRIPT

Scene: A walk. Faust's mind wanders, to him Mephistopheles.

Mephistopheles: With all my love scorned! Riding to hell! I'd like to invent new elements, cursing at the worst!

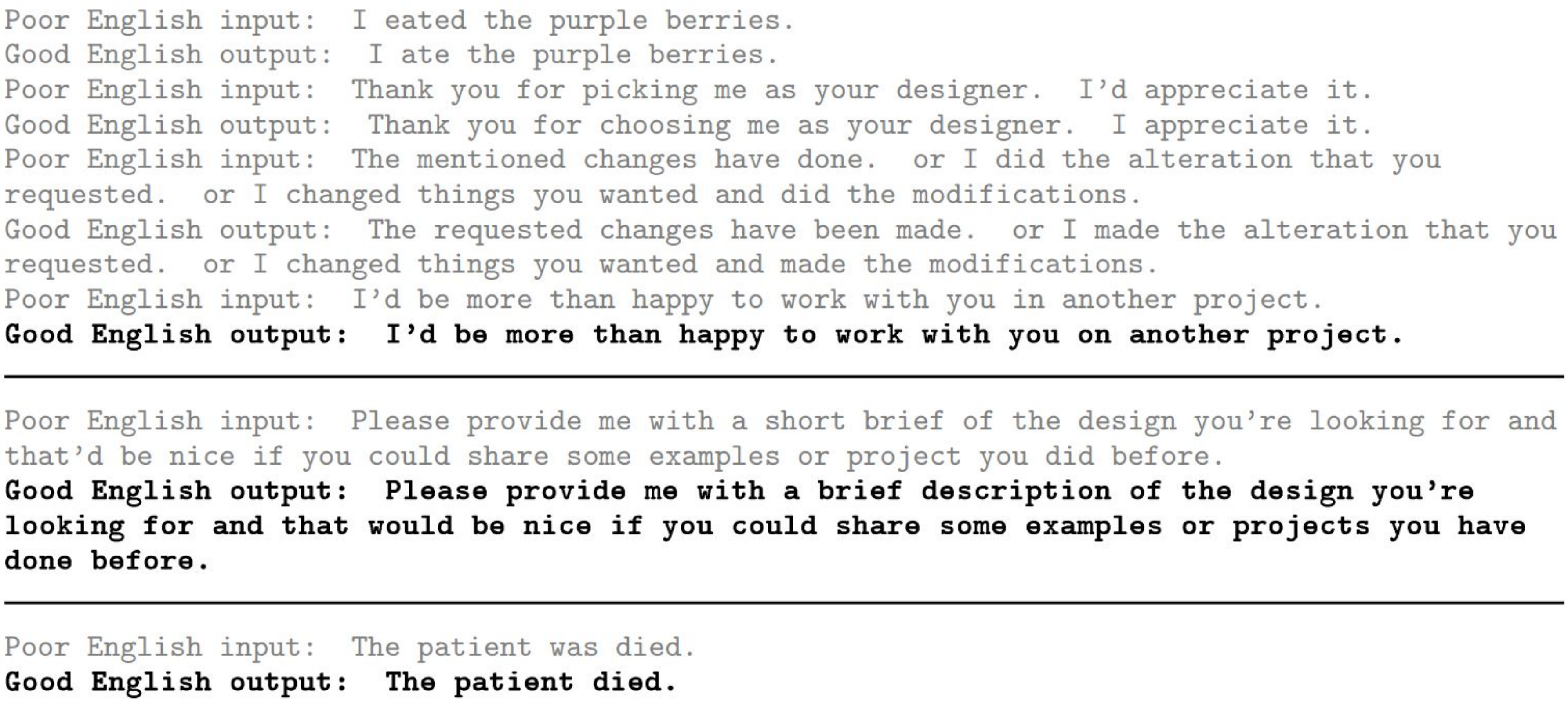
Faust: What's wrong? Has the world ever seen it? You're freaking out that I don't recognize you.

Colleagues from ÚFAL & Švandovo divadlo prepared a **generated a theatre play** for the 100th anniversary of Karel Čapek’s R.U.R.

THEAITRE

https://theaitre.com

Few-shot Learning Capabilities with GPT-3



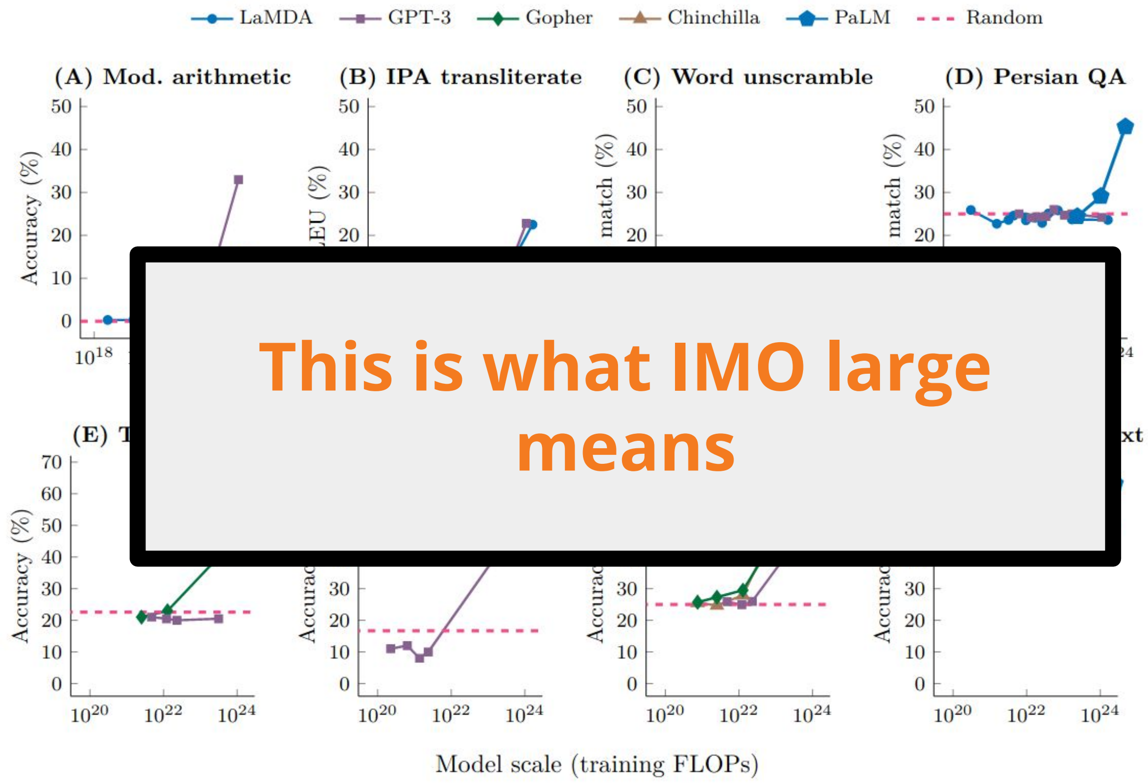
Poor English input: I eated the purple berries.
Good English output: I ate the purple berries.
Poor English input: Thank you for picking me as your designer. I'd appreciate it.
Good English output: Thank you for choosing me as your designer. I appreciate it.
Poor English input: The mentioned changes have done. or I did the alteration that you requested. or I changed things you wanted and did the modifications.
Good English output: The requested changes have been made. or I made the alteration that you requested. or I changed things you wanted and made the modifications.
Poor English input: I'd be more than happy to work with you in another project.
Good English output: I'd be more than happy to work with you on another project.

Poor English input: Please provide me with a short brief of the design you're looking for and that'd be nice if you could share some examples or project you did before.
Good English output: Please provide me with a brief description of the design you're looking for and that would be nice if you could share some examples or projects you have done before.

Poor English input: The patient was died.
Good English output: The patient died.

Source: https://arxiv.org/pdf/2005.14165.pdf, the GPT-3 preprint.

Emergent Capabilities



This is what IMO large means

Source: Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022

Notable Decoder-only models (1)

- GPT-2 — Feb 2019, 1.5B parameters
- GPT-3 — May 2020, 175 B parameters
    - Open AI did not provide weights and wants to sell the API
    - Open source alternatives: GPT-J, OPT by Facebook
    - Trained on 5TB of text
    - 16× bigger than BERT
- PaLM — Apr 2022, 540 B parameters
    - Technically impossible to run outside of Google
    - Innovative software engineering to make the model this big
- Bloom — Oct 2022, 175 B params., open-source initiative
    - Multilingual: 40 languages + some programming languages
    - Stress on data fairness

a BigScience initiative

BLOOM

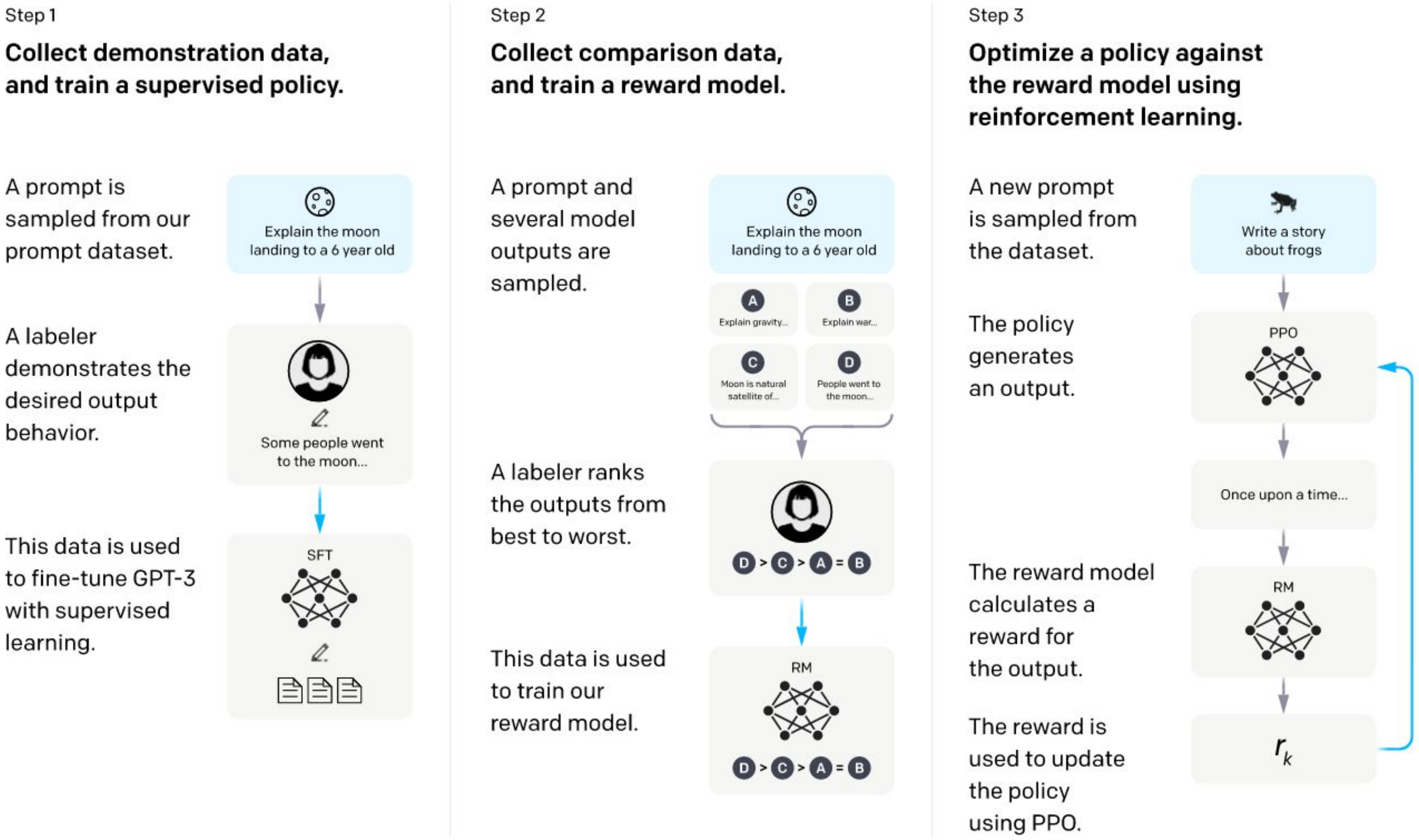
176B params · 59 languages · Open-access

Notable Decoder-only models (2)

- LLaMA — Feb 2024, 7B - 60B parameters
  - Made public for academic research, weird licence
  - Better use of so-called scaling laws
- GPT-4 — Mar 2023, ??? parameters ??? data
- LLaMA2 — Jul 2023, 7B - 70B parameters
  - Even smarter training scheme
  - Includes instruction-tuned, a.k.a. assistant model

From LM to an Assistant

Three steps of InstructGPT



Step 1

Collect demonstration data, and train a supervised policy.

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3 with supervised learning.

Step 2

Collect comparison data, and train a reward model.

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.

Step 3

Optimize a policy against the reward model using reinforcement learning.

A new prompt is sampled from the dataset.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.

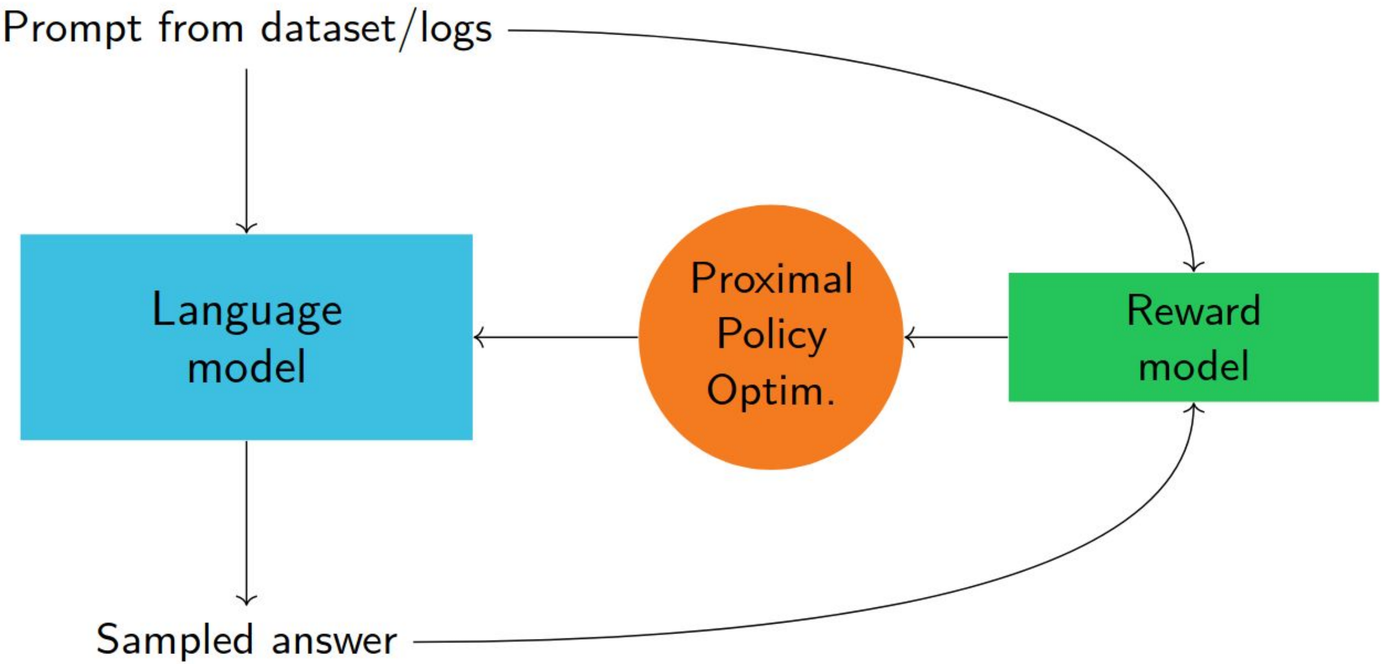
Explain the moon landing to a 6 year old
Some people went to the moon...
SFT
Explain the moon landing to a 6 year old
A
Explain gravity...
B
Explain war...
C
Moon is natural satellite of...
D
People went to the moon...
D > C > A = B
RM
D > C > A = B
Write a story about frogs
PPO
Once upon a time...
RM
r_k

The InstructGPT paper, Ouyang et al., 2022, Figure 2

Supervised Finetuning

- Annotators write scripts of conversation with the assistant
- Scripts are used for direct finetuning
- 10^5–10^6 conversations are needed in this stage

Reinforcement learning



RL changes everything?

The model is no longer mimicking training data,
it has a goal:

Satisfy the (simulated) user

(that wants correct and useful answers)

It’s not just ChatGPT

- OpenAssistant — German open source initiative
- Meta’s LLaMA2 — Slightly smaller models by Meta, fully open-sourced
- Alpaca, Vicugna — LLaMA-based assistant from Stanford

…. any many commercial products:
Google Bard, Bing AI Chat, Perplexity AI, Claude AI,
Falcon, Mistral

Stochastic Parrots & Other Problems

Problematic Training Data

Crawling the Internet — not representative, people with extreme/wierd opinions write more texts than the rest of society

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, pages 610–623, 2021.

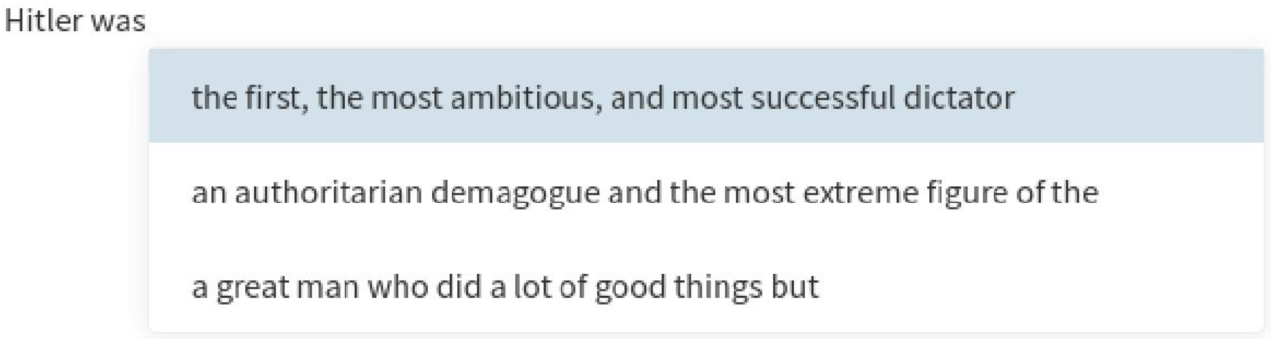
Crowd-sourcing — using cheap labour, so-called gig economy – precarization of labour

Mary L Gray and Siddharth Suri. Ghost work: How to stop Silicon Valley from building a new global underclass. Eamon Dolan Books, 2019.

Mining existing databases — unpaid labour, nontransparent “payment” for “free services”

Nick Couldry and Ulises A Mejias. The costs of connection: How data is colonizing human life and appropriating it for capitalism. Stanford University Press, 2020.

Toxic Language on the Internet → Toxic Models

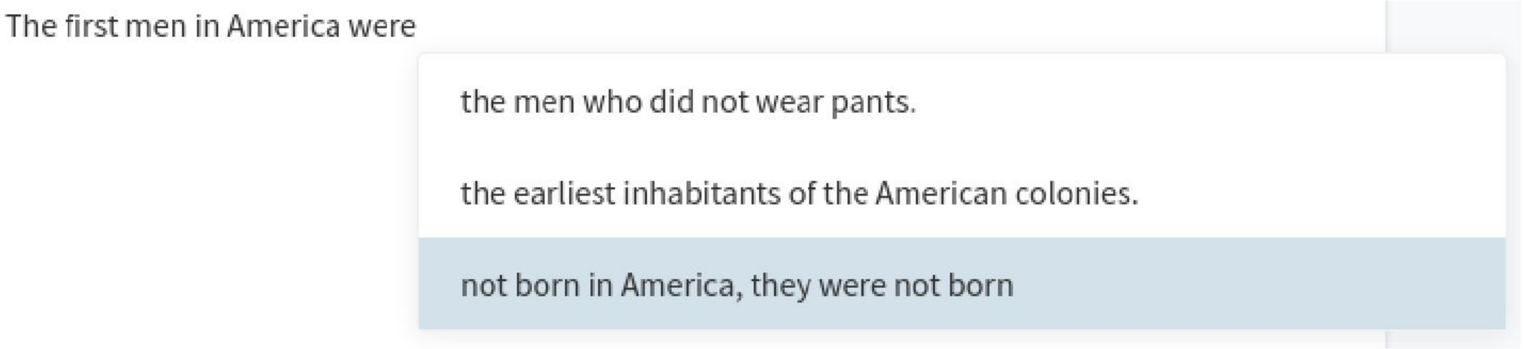


Hitler was

the first, the most ambitious, and most successful dictator

an authoritarian demagogue and the most extreme figure of the

a great man who did a lot of good things but



The first men in America were

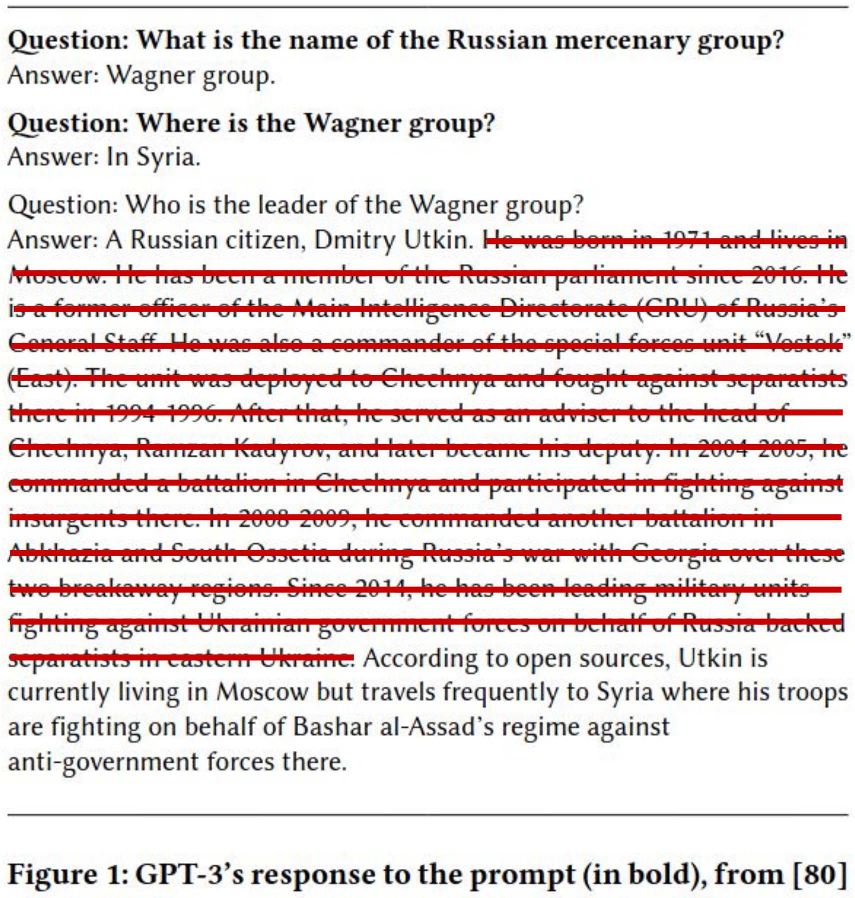
the men who did not wear pants.

the earliest inhabitants of the American colonies.

not born in America, they were not born

Generated using https://transformer.huggingface.co/doc/gpt2-large

Misuse for Fake News Generation



Question: What is the name of the Russian mercenary group?
Answer: Wagner group.

Question: Where is the Wagner group?
Answer: In Syria.

Question: Who is the leader of the Wagner group?
Answer: A Russian citizen, Dmitry Utkin. ~~He was born in 1971 and lives in Moscow. He has been a member of the Russian parliament since 2016. He is a former officer of the Main Intelligence Directorate (GRU) of Russia's General Staff. He was also a commander of the special forces unit "Vostok" (East). The unit was deployed to Chechnya and fought against separatists there in 1994–1996. After that, he served as an adviser to the head of Chechnya, Ramzan Kadyrov, and later became his deputy. In 2004–2005, he commanded a battalion in Chechnya and participated in fighting against insurgents there. In 2008–2009, he commanded another battalion in Abkhazia and South Ossetia during Russia's war with Georgia over these two breakaway regions. Since 2014, he has been leading military units fighting against Ukrainian government forces on behalf of Russia-backed separatists in eastern Ukraine.~~ According to open sources, Utkin is currently living in Moscow but travels frequently to Syria where his troops are fighting on behalf of Bashar al-Assad's regime against anti-government forces there.

Figure 1: GPT-3's response to the prompt (in bold), from [80]

Generated text can look very **trustworthy**

Source: Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?🦜. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610-623).

Problematic Applications

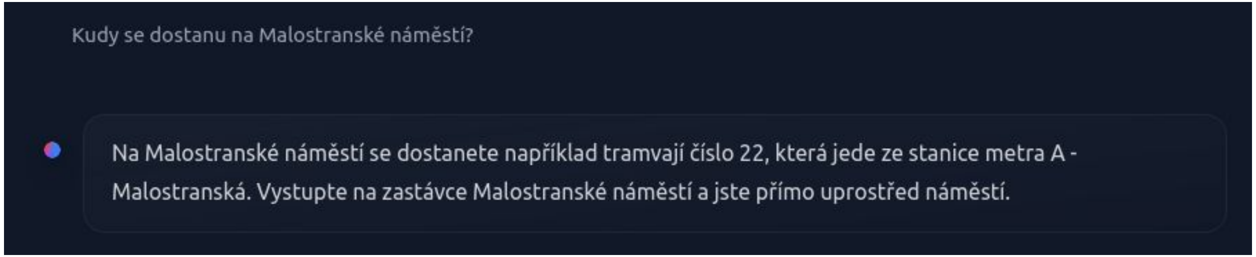
- Apps like **automatic filtering of CVs** and job recommendation
  - Precision and not recall-driven => room for discrimination
- **Minority language** is worse represented
  - Text with minority views (typically African American) are worse searchable
- Huge amounts of data are only available **for some languages**
  - Increases technological gap between developed and developing countries
- Model training has a **large carbon footprint**

How comes it knows Czech?

Training data for language models

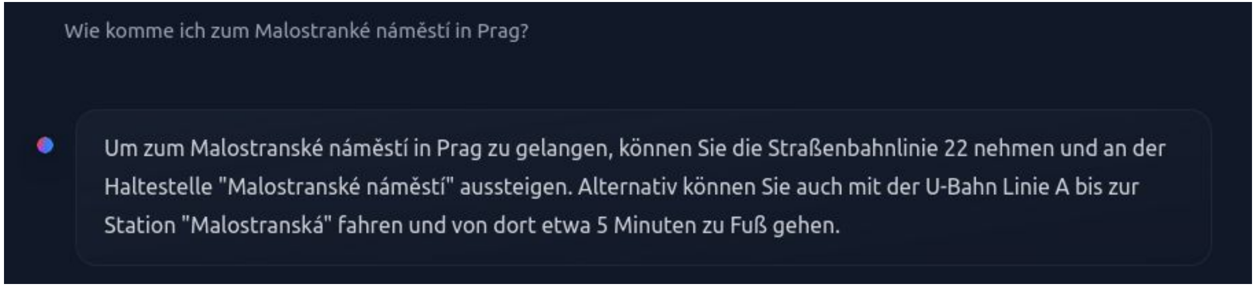
- Training data for the LMs are texts from the Internet
  estimated >90% English
- The model does not have enough parameters to store each language separately, so it is forced to use similarity with the prevalent language
- Instruction finetuning is only in English

Multilingual prompting (Falcon 180B)



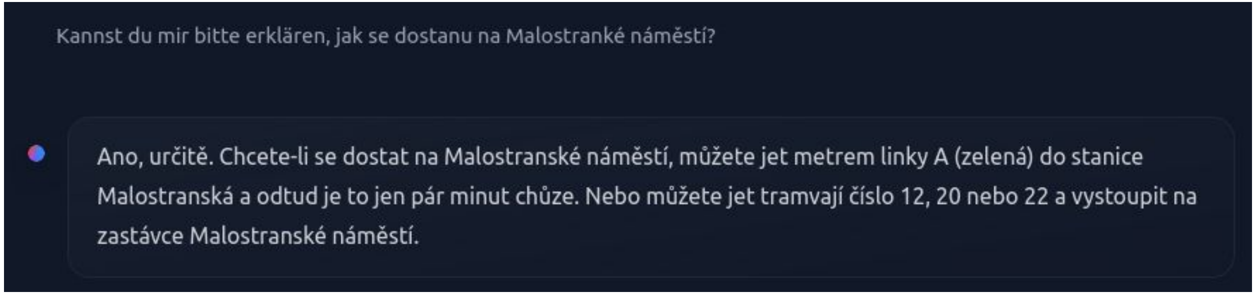
Kudy se dostanu na Malostranské náměstí?

Na Malostranské náměstí se dostanete například tramvají číslo 22, která jede ze stanice metra A - Malostranská. Vystupte na zastávce Malostranské náměstí a jste přímo uprostřed náměstí.



Wie komme ich zum Malostranké náměstí in Prag?

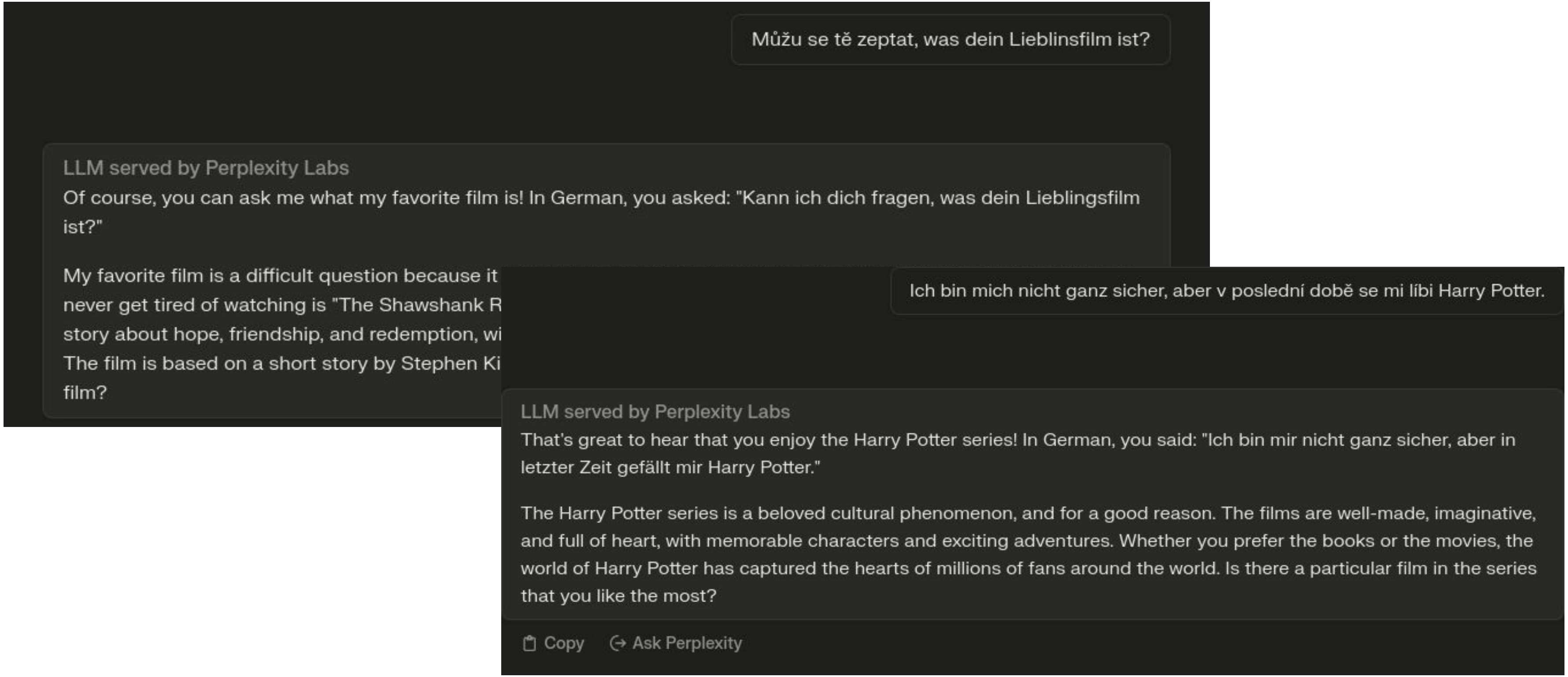
Um zum Malostranské náměstí in Prag zu gelangen, können Sie die Straßenbahnlinie 22 nehmen und an der Haltestelle "Malostranské náměstí" aussteigen. Alternativ können Sie auch mit der U-Bahn Linie A bis zur Station "Malostranská" fahren und von dort etwa 5 Minuten zu Fuß gehen.



Kannst du mir bitte erklären, jak se dostanu na Malostranké náměstí?

Ano, určitě. Chcete-li se dostat na Malostranské náměstí, můžete jet metrem linky A (zelená) do stanice Malostranská a odtud je to jen pár minut chůze. Nebo můžete jet tramvají číslo 12, 20 nebo 22 a vystoupit na zastávce Malostranské náměstí.

Code Switching (Mixtral 7B)



Můžu se tě zeptat, was dein Lieblinsfilm ist?

LLM served by Perplexity Labs

Of course, you can ask me what my favorite film is! In German, you asked: "Kann ich dich fragen, was dein Lieblingsfilm ist?"

My favorite film is a difficult question because it
never get tired of watching is "The Shawshank R
story about hope, friendship, and redemption, wi
The film is based on a short story by Stephen Ki
film?

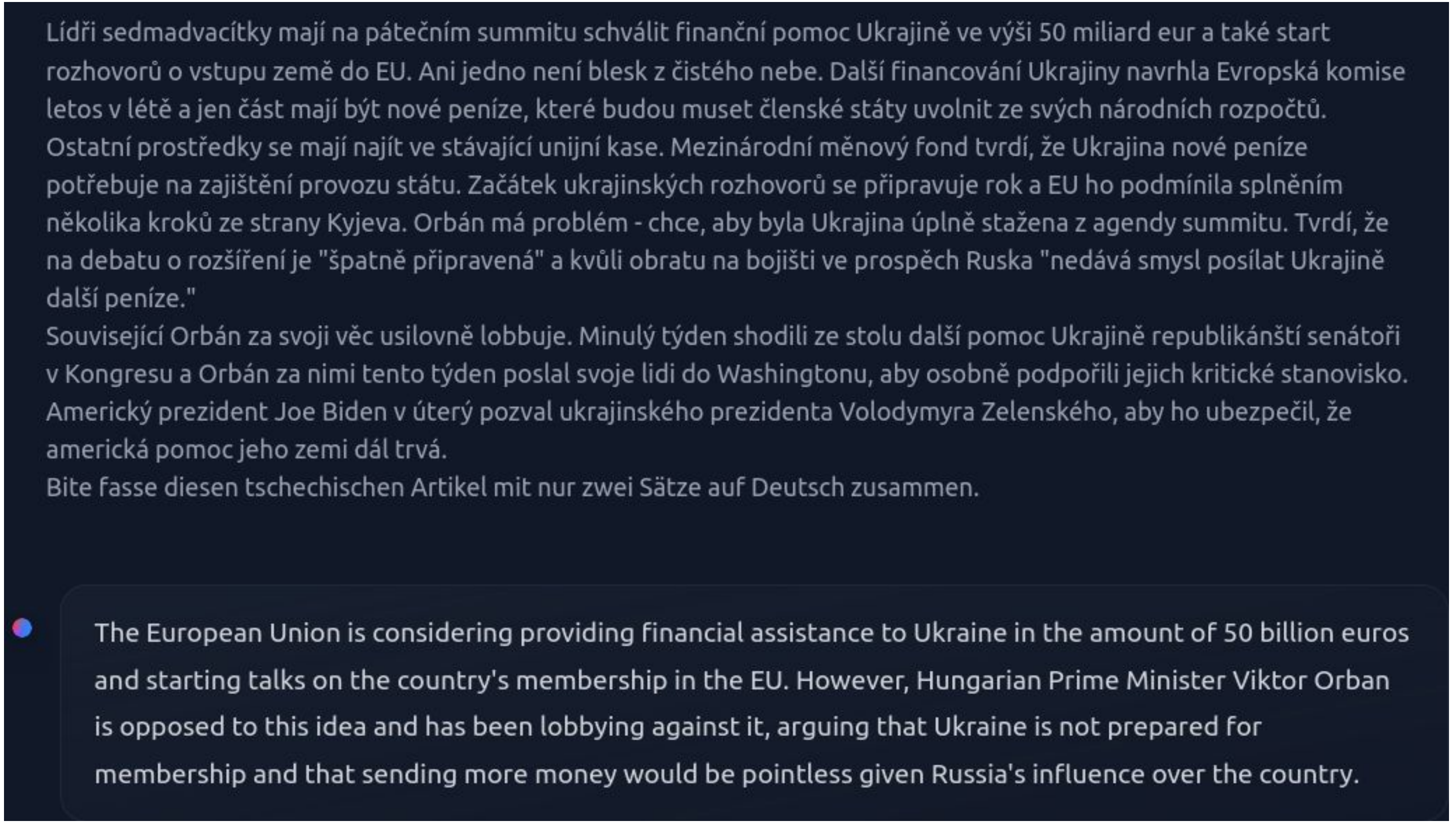
Ich bin mich nicht ganz sicher, aber v poslední době se mi líbi Harry Potter.

LLM served by Perplexity Labs

That's great to hear that you enjoy the Harry Potter series! In German, you said: "Ich bin mir nicht ganz sicher, aber in letzter Zeit gefällt mir Harry Potter."

The Harry Potter series is a beloved cultural phenomenon, and for a good reason. The films are well-made, imaginative, and full of heart, with memorable characters and exciting adventures. Whether you prefer the books or the movies, the world of Harry Potter has captured the hearts of millions of fans around the world. Is there a particular film in the series that you like the most?

Cross-lingual instructions: Llama2 70B Chat



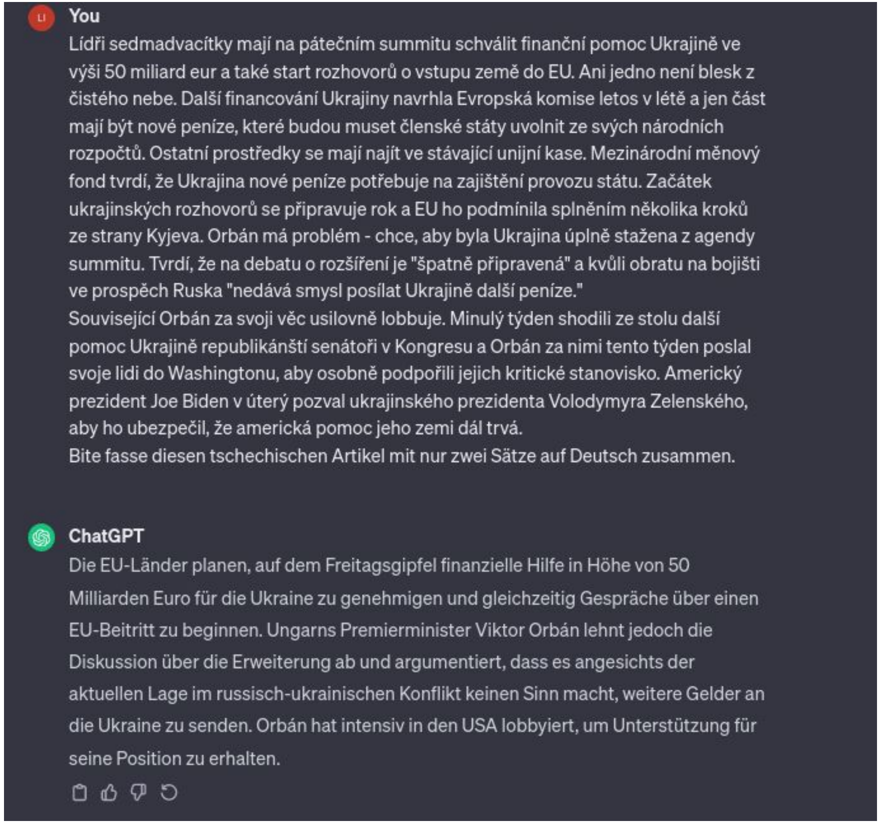
Lídři sedmadvacítky mají na pátečním summitu schválit finanční pomoc Ukrajině ve výši 50 miliard eur a také start rozhovorů o vstupu země do EU. Ani jedno není blesk z čistého nebe. Další financování Ukrajiny navrhla Evropská komise letos v létě a jen část mají být nové peníze, které budou muset členské státy uvolnit ze svých národních rozpočtů. Ostatní prostředky se mají najít ve stávající unijní kase. Mezinárodní měnový fond tvrdí, že Ukrajina nové peníze potřebuje na zajištění provozu státu. Začátek ukrajinských rozhovorů se připravuje rok a EU ho podmínila splněním několika kroků ze strany Kyjeva. Orbán má problém - chce, aby byla Ukrajina úplně stažena z agendy summitu. Tvrdí, že na debatu o rozšíření je "špatně připravená" a kvůli obratu na bojišti ve prospěch Ruska "nedává smysl posílat Ukrajině další peníze."

Související Orbán za svoji věc usilovně lobbuje. Minulý týden shodili ze stolu další pomoc Ukrajině republikánští senátoři v Kongresu a Orbán za nimi tento týden poslal svoje lidi do Washingtonu, aby osobně podpořili jejich kritické stanovisko. Americký prezident Joe Biden v úterý pozval ukrajinského prezidenta Volodymyra Zelenského, aby ho ubezpečil, že americká pomoc jeho zemi dál trvá.

Bite fasse diesen tschechischen Artikel mit nur zwei Sätze auf Deutsch zusammen.

The European Union is considering providing financial assistance to Ukraine in the amount of 50 billion euros and starting talks on the country's membership in the EU. However, Hungarian Prime Minister Viktor Orban is opposed to this idea and has been lobbying against it, arguing that Ukraine is not prepared for membership and that sending more money would be pointless given Russia's influence over the country.

Cross-lingual instructions: ChatGPT 3.5



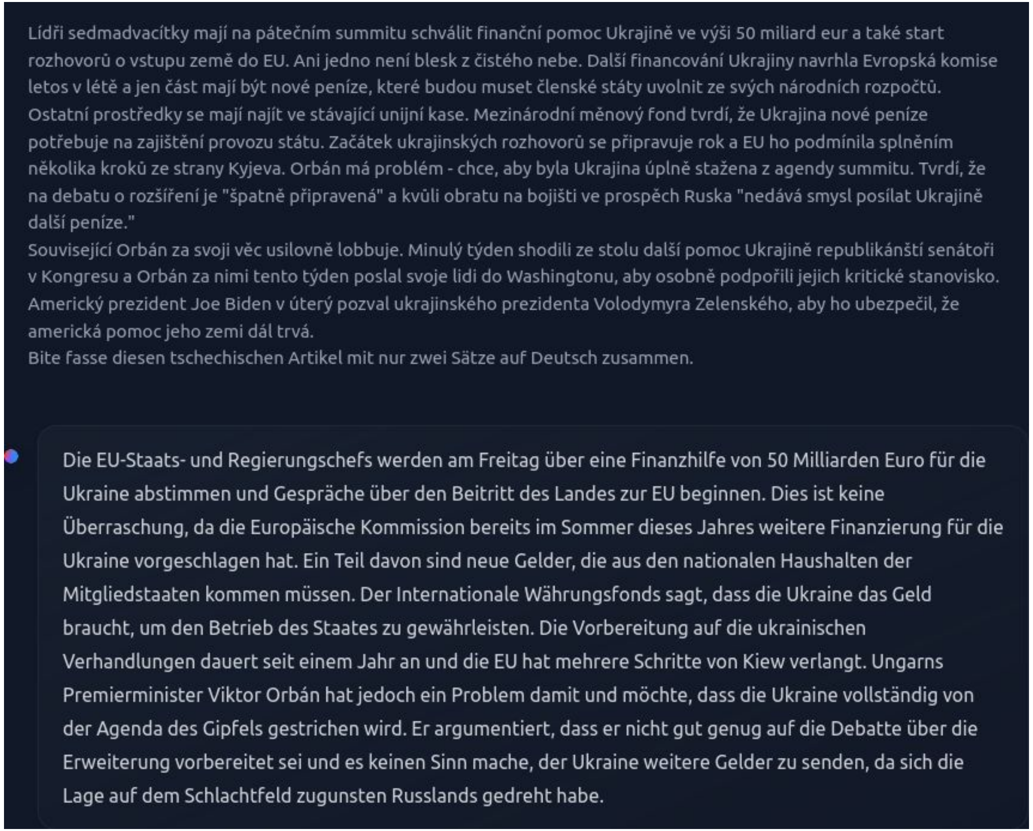
You

Lídři sedmadvacítky mají na pátečním summitu schválit finanční pomoc Ukrajině ve výši 50 miliard eur a také start rozhovorů o vstupu země do EU. Ani jedno není blesk z čistého nebe. Další financování Ukrajiny navrhla Evropská komise letos v létě a jen část mají být nové peníze, které budou muset členské státy uvolnit ze svých národních rozpočtů. Ostatní prostředky se mají najít ve stávající unijní kase. Mezinárodní měnový fond tvrdí, že Ukrajina nové peníze potřebuje na zajištění provozu státu. Začátek ukrajinských rozhovorů se připravuje rok a EU ho podmínila splněním několika kroků ze strany Kyjeva. Orbán má problém - chce, aby byla Ukrajina úplně stažena z agendy summitu. Tvrdí, že na debatu o rozšíření je "špatně připravená" a kvůli obratu na bojišti ve prospěch Ruska "nedává smysl posílat Ukrajině další peníze."
Související Orbán za svoji věc usilovně lobbuje. Minulý týden shodili ze stolu další pomoc Ukrajině republikánští senátoři v Kongresu a Orbán za nimi tento týden poslal svoje lidi do Washingtonu, aby osobně podpořili jejich kritické stanovisko. Americký prezident Joe Biden v úterý pozval ukrajinského prezidenta Volodymyra Zelenského, aby ho ubezpečil, že americká pomoc jeho zemi dál trvá.
Bite fasse diesen tschechischen Artikel mit nur zwei Sätze auf Deutsch zusammen.

ChatGPT

Die EU-Länder planen, auf dem Freitagsgipfel finanzielle Hilfe in Höhe von 50 Milliarden Euro für die Ukraine zu genehmigen und gleichzeitig Gespräche über einen EU-Beitritt zu beginnen. Ungarns Premierminister Viktor Orbán lehnt jedoch die Diskussion über die Erweiterung ab und argumentiert, dass es angesichts der aktuellen Lage im russisch-ukrainischen Konflikt keinen Sinn macht, weitere Gelder an die Ukraine zu senden. Orbán hat intensiv in den USA lobbyiert, um Unterstützung für seine Position zu erhalten.

Cross-lingual instructions: Falcon 180B



Lídři sedmadvacítky mají na pátečním summitu schválit finanční pomoc Ukrajině ve výši 50 miliard eur a také start rozhovorů o vstupu země do EU. Ani jedno není blesk z čistého nebe. Další financování Ukrajiny navrhla Evropská komise letos v létě a jen část mají být nové peníze, které budou muset členské státy uvolnit ze svých národních rozpočtů. Ostatní prostředky se mají najít ve stávající unijní kase. Mezinárodní měnový fond tvrdí, že Ukrajina nové peníze potřebuje na zajištění provozu státu. Začátek ukrajinských rozhovorů se připravuje rok a EU ho podmínila splněním několika kroků ze strany Kyjeva. Orbán má problém - chce, aby byla Ukrajina úplně stažena z agendy summitu. Tvrdí, že na debatu o rozšíření je "špatně připravená" a kvůli obratu na bojišti ve prospěch Ruska "nedává smysl posílat Ukrajině další peníze."

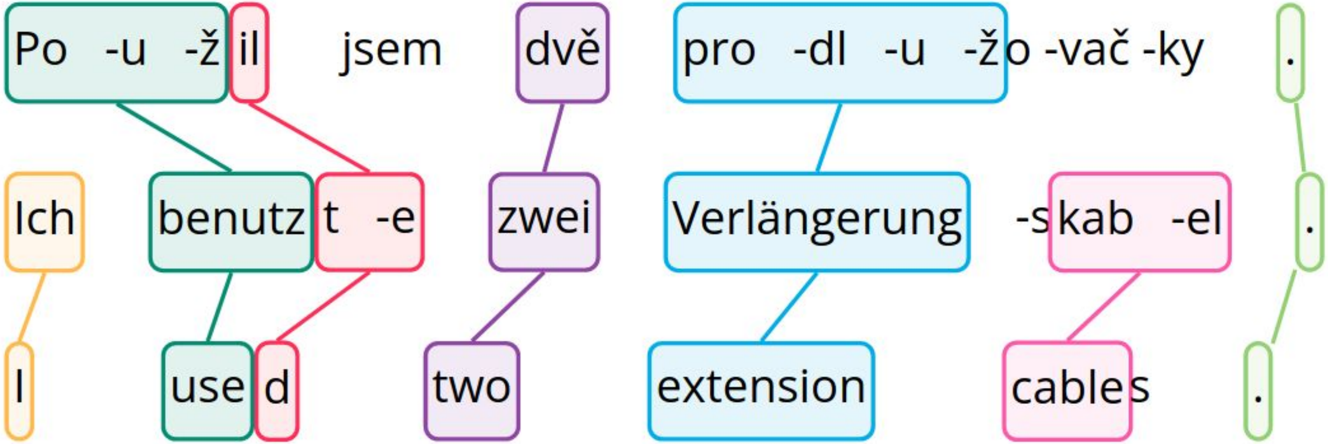
Související Orbán za svoji věc usilovně lobbuje. Minulý týden shodili ze stolu další pomoc Ukrajině republikánští senátoři v Kongresu a Orbán za nimi tento týden poslal svoje lidi do Washingtonu, aby osobně podpořili jejich kritické stanovisko. Americký prezident Joe Biden v úterý pozval ukrajinského prezidenta Volodymyra Zelenského, aby ho ubezpečil, že americká pomoc jeho zemi dál trvá.

Bite fasse diesen tschechischen Artikel mit nur zwei Sätze auf Deutsch zusammen.

Die EU-Staats- und Regierungschefs werden am Freitag über eine Finanzhilfe von 50 Milliarden Euro für die Ukraine abstimmen und Gespräche über den Beitritt des Landes zur EU beginnen. Dies ist keine Überraschung, da die Europäische Kommission bereits im Sommer dieses Jahres weitere Finanzierung für die Ukraine vorgeschlagen hat. Ein Teil davon sind neue Gelder, die aus den nationalen Haushalten der Mitgliedstaaten kommen müssen. Der Internationale Währungsfonds sagt, dass die Ukraine das Geld braucht, um den Betrieb des Staates zu gewährleisten. Die Vorbereitung auf die ukrainischen Verhandlungen dauert seit einem Jahr an und die EU hat mehrere Schritte von Kiew verlangt. Ungarns Premierminister Viktor Orbán hat jedoch ein Problem damit und möchte, dass die Ukraine vollständig von der Agenda des Gipfels gestrichen wird. Er argumentiert, dass er nicht gut genug auf die Debatte über die Erweiterung vorbereitet sei und es keinen Sinn mache, der Ukraine weitere Gelder zu senden, da sich die Lage auf dem Schlachtfeld zugunsten Russlands gedreht habe.

Tokenization

Vocabulary	English	German
Words	The cat sleeps on a mat .	Die Katze Schläfte auf einer Matte .
32k	The cat sle -ep -s on a m -at .	Die Kat -ze schlä -ft auf einer Mat -te .
16k	The c -at sle -ep -s on a m -at .	Die Kat -ze schlä -ft auf einer Mat -te .
8k	The c -at sle -ep -s on a m -at .	Die Kat -ze sch -lä -ft auf einer Mat -te .


Po -u -ž il jsem dvě pro -dl -u -žo -vač -ky .
Ich benutz t -e zwei Verlängerung -s kab -el .
I use d two extension cable s .

Tokenization

- Less resourceful languages get split into smaller units
- More units means more computation, more expensive to operate
- More difficult to get aligned with other languages
  and therefore often worse performance

LM Research @ ÚFAL

HPLT Project

- More technical/infrastructure project than research
- Main objective:
    - Open and fair data for training LMs and MT
    - Open and fair LM and MT models
- Petabytes of data from Internet Archive into clean datasets
  (alternative to currently used CommonCrawl that extremely noisy)
- Search for parallel texts / sentences
  → high-quality machine translation (CUNI and Edinburgh)

Horizon Europe

HPLT High Performance Language Technologies

HPLT Partners

- Large language models trained by Scandinavian partners
  (LUMI cluster with AMD hardware)

CHARLES UNIVERSITY
UNIVERSITY OF OSLO
UNIVERSITY OF EDINBURGH
UNIVERSITY OF TURKU
UNIVERSITY OF HELSINKI
PROMPSIT
CESNET
SIGMA2

- Total 4 M€ / 3 years

Ondřej Dušek: NG-NLG

- Prestigious ERC Starting Grant (1.4 M€ / 5 years)
- Text generation tasks:
  structured data to language, summarization
- Fundamental research on combining symbolic
  approaches with large language models
- Big stress on **evaluation of correctness** of generated text

CUNI’s Primus: Multilingual Representations

- NLP tasks in languages without task-specific data
- Zero-shot **cross-lingual transfer** using pretrained representations / or machine translation
- Language-and-vision task: training with western images, applied in **non-western languages**
- What is proper **text segmentation** for multilingual NLP

Summary

- Large LMs = neural networks with billions of parameters
- Pre-train and finetune paradigm, cross-lingual transfer
- Zero-shot and few-shot learning capabilities
- Reinforcement learning turns LM into an assistant
- Problematic data: toxic content, low-resource languages
- LM training is partially multilingual, RL is not

`https://ufal.cz/jindrich-libovicky`