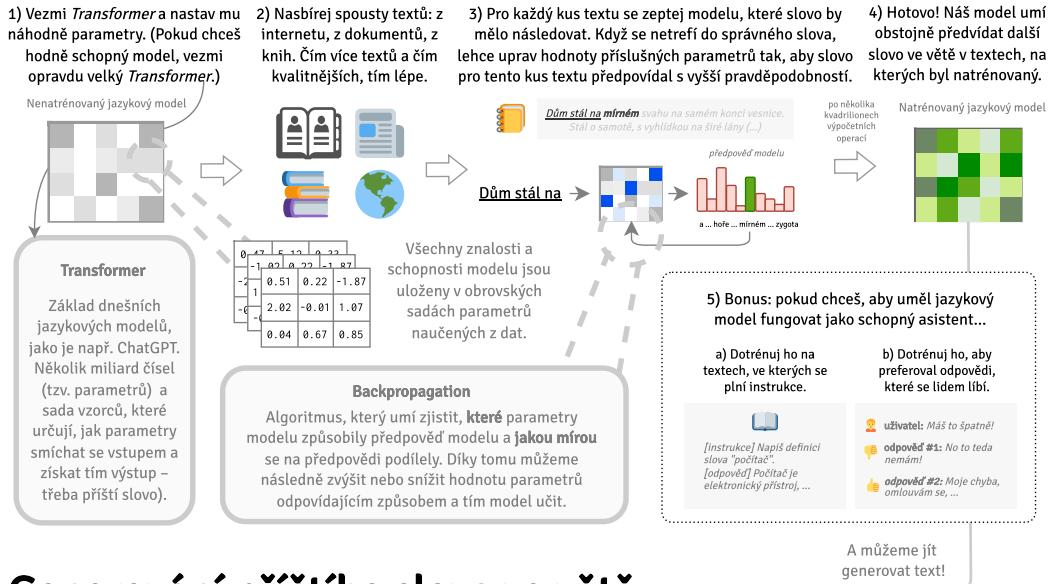
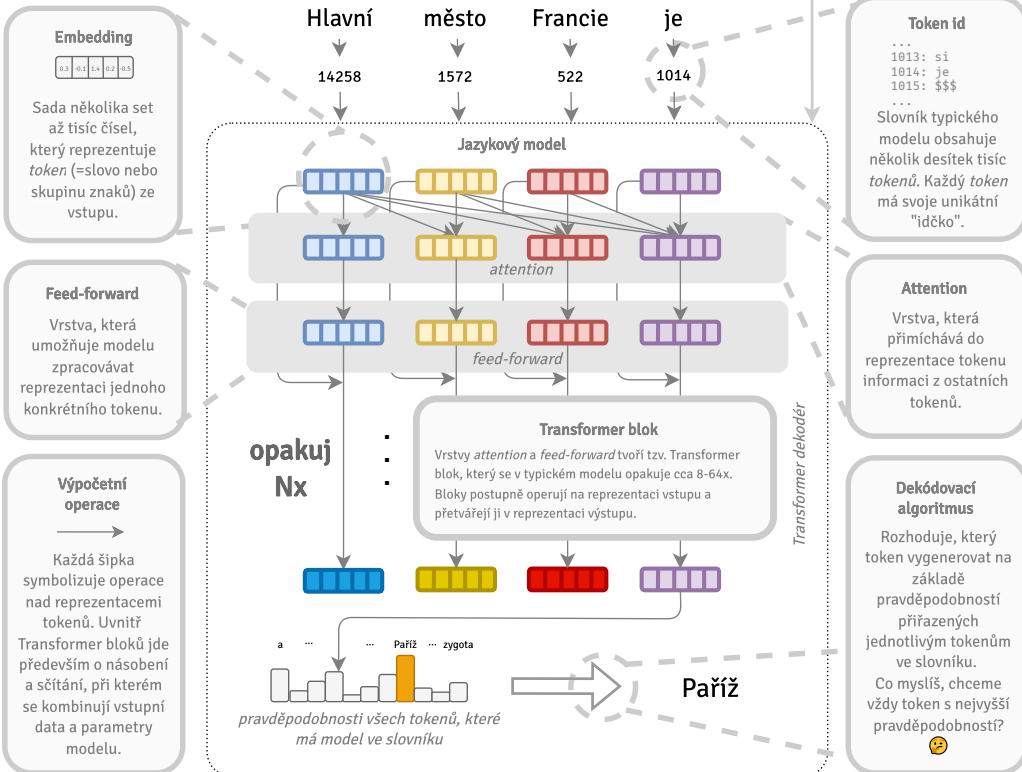


Recept na model, který pohání ChatGPT



Generování příštího slova ve větě



NOVĚ: Toto vše najdete animované na <https://animatedllm.github.io/>!

Generování textu: otázky a odpovědi

Umí si jazykové modely i hledat v internetu?



Krátká odpověď: Neumí, ale někdo to může dělat za ně.

Dlouhá odpověď: Jazykový model jako takový hledat na internetu neumí. Všechno, co umí, se musel naučit během trénování.

V dnešní době ale mají komerční jazykové modely typicky externí vyhledávací modul. Tento modul hledá informace související s dotazem a předkládá je jazykovému modelu na vstupu společně se vstupem od uživatele. Model pak s nimi pracuje stejně, jako se zbytkem textu. Kvalita těchto výsledků může ovlivnit vygenerovaný text jak pozitivně, tak negativně.

Má model seznam všech slov ve všech jazycech?



Krátká odpověď: Ne – ale jejich částí ano!

Dlouhá odpověď: Všechna slova bylo opravdu až příliš (už jen všechny vyskloňované tvary slov v češtině!). Proto používáme jako tokeny tzv. "subwordy": slova a části slov, se kterými můžeme libovolná slova poskládat. Frekventovanější slova máme ve slovníku přímo, ta méně častá složíme z více částí.

Rozsekáme text na subwordy a pak ho zase poskládat je úkolem specializovaného algoritmu, tzv. tokenizéra, který pracuje nezávisle na jazykovém modelu.

Jak energeticky náročné je natřenovat model? A kolik energie spotřebuje jeden dotaz?

Krátká odpověď: Trénovat je náročné, generovat ne tolik.

Dlouhá odpověď: Záleží na velikosti modelu, ale odhaduje se, že natřenovat model s 175 mld. parametry stojí 1.2 GWh energie, což odpovídá roční spotřebě 120 amerických domácností. Modely se ale naštěstí netřenují tak často.

Generovat na natřenovaném modelu text už je mnohem efektivnější: průměrná odpověď ChatGPT spotřebuje cca 0,3 Wh, což odpovídá např. rychlovárně konvici zapnuté na 1 sekundu.

[1] <https://blog.camniran.com/the-point-of-singularity>
[2] <https://idioti.org/how-much-energy-do-lms-consume-unveiling-the-power-behind-ai/>
[3] <https://andymasley.substack.com/p/a-cheat-sheet-for-conversations-about>

Co dělá velké jazykové modely tak inteligentní?



Krátká odpověď: To je mi ale těžká otázka...!

Dlouhá odpověď: Modely do nějaké míry umějí zopakovat to, co viděly v trénovacích datech. Chytrá odpověď proto mohla jen "ležet na internetu".

Modely ale umí částečně i generalizovat: kombinovat naučené vzory v originální odpovědi. Důvodem může být to, že s velkým množstvím dat může být generalizovat jednodušší, než se učit vše nazepamět.

A tady už přichází další otázky: Co dělá inteligentní člověka? A co je to vlastně ta "intelligence"?

Kde se můžu naučit o jazykových modelech více?

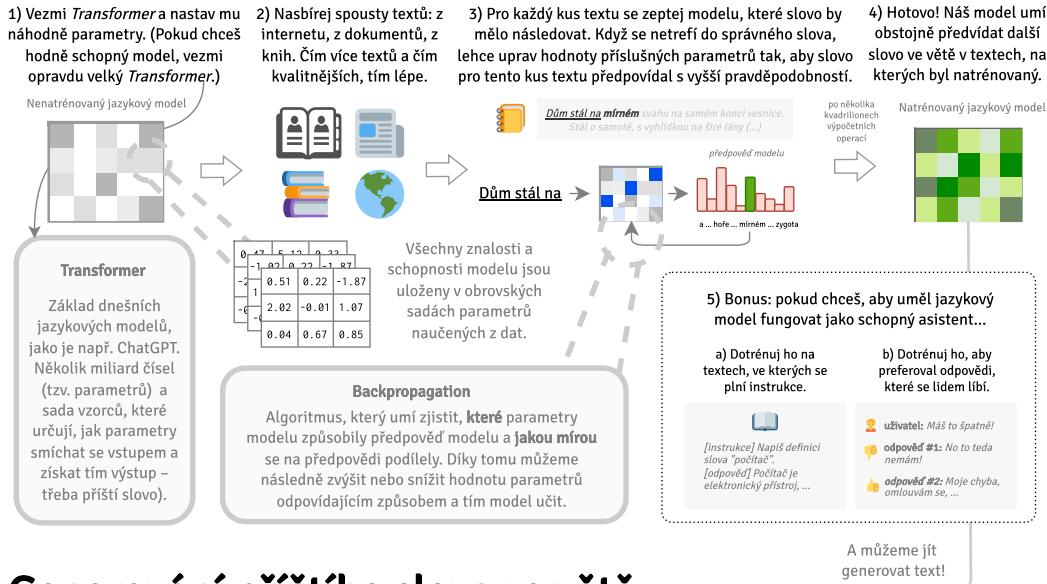


Krátká odpověď: U nás na ÚFALu!

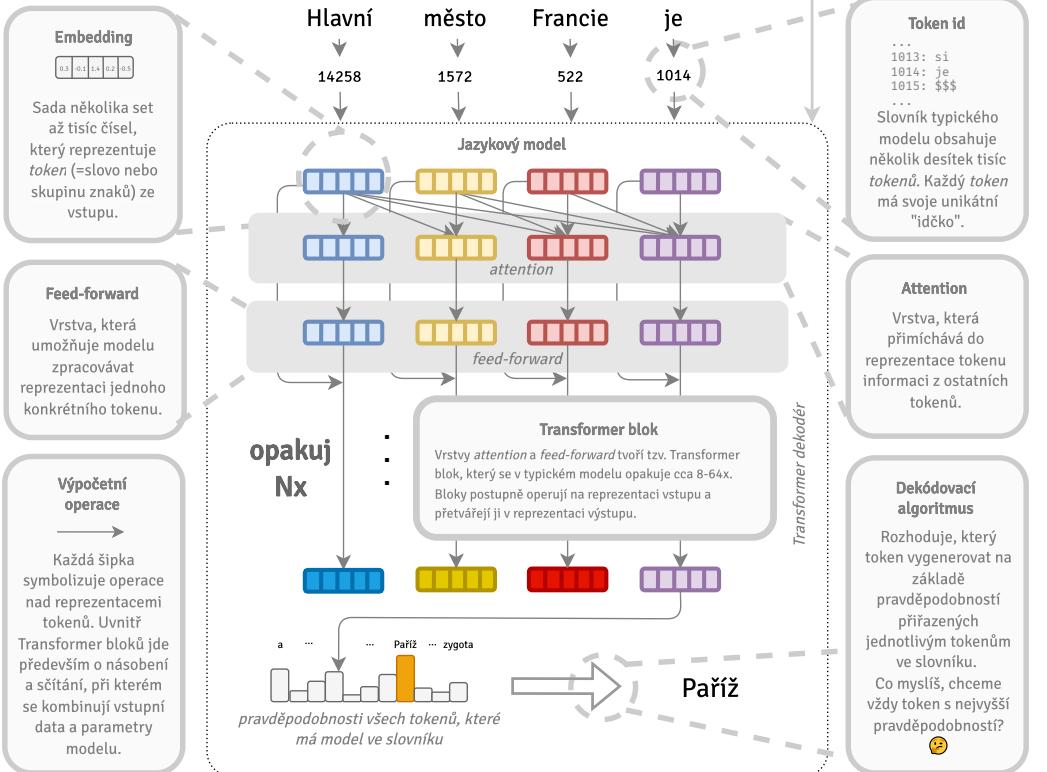
Dlouhá odpověď: Podobnými tématy se zabýváme na Ústavu Formální a Aplikované Lingvistiky (ÚFAL), což je součást Matematicko-fyzikální fakulty Univerzity Karlové.

Na naší katedře propoujeme znalosti o jazyce se znalostmi ze strojového učení. Učíme o tom bakalářské i magisterské předměty, jako třeba Deep Learning nebo Large Language Models.

Recept na model, který pohání ChatGPT



Generování příštího slova ve větě



NOVĚ: Toto vše najdete animované na <https://animatedllm.github.io/>!

Generování textu: otázky a odpovědi

Umí si jazykové modely i hledat v internetu?



Krátká odpověď: Neumí, ale někdo to může dělat za ně.

Dlouhá odpověď: Jazykový model jako takový hledat na internetu neumí. Všechno, co umí, se musel naučit během trénování.

V dnešní době ale mají komerční jazykové modely typicky externí vyhledávací modul. Tento modul hledá informace související s dotazem a předkládá je jazykovému modelu na vstupu společně se vstupem od uživatele. Model pak s nimi pracuje stejně, jako se zbytkem textu. Kvalita těchto výsledků může ovlivnit vygenerovaný text jak pozitivně, tak negativně.

Má model seznam všech slov ve všech jazycech?



Krátká odpověď: Ne – ale jejich částí ano!

Dlouhá odpověď: Všechna slova bylo opravdu až příliš (už jen všechny vyskloňované tvary slov v češtině!). Proto používáme jako tokeny tzv. "subwordy": slova a části slov, se kterými můžeme libovolná slova poskládat. Frekventovanější slova máme ve slovníku přímo, ta méně častá složíme z více částí.

Rozsekáme text na subwordy a pak ho zase poskládat je úkolem specializovaného algoritmu, tzv. tokenizéra, který pracuje nezávisle na jazykovém modelu.

Jak energeticky náročné je natřenovat model? A kolik energie spotřebuje jeden dotaz?

Krátká odpověď: Trénovat je náročné, generovat ne tak.

Dlouhá odpověď: Záleží na velikosti modelu, ale odhaduje se, že natřenovat model s 175 mld. parametry stojí 1.2 GWh energie, což odpovídá roční spotřebě 120 amerických domácností. Modely se ale naštěstí netřenují tak často.

Generovat na natřenovaném modelu text už je mnohem efektivnější: průměrná odpověď ChatGPT spotřebuje cca 0,3 Wh, což odpovídá např. rychlovárně konvici zapnuté na 1 sekundu.

[1] <https://blog.camniran.com/the-point-of-singularity>
[2] <https://idioti.org/how-much-energy-do-lms-consume-unveiling-the-power-behind-ai/>
[3] <https://andymasley.substack.com/p/a-cheat-sheet-for-conversations-about>

Co dělá velké jazykové modely tak inteligentní?



Krátká odpověď: To je mi ale těžká otázka...!

Dlouhá odpověď: Modely do nějaké míry umějí zopakovat to, co viděly v trénovacích datech. Chytrá odpověď proto mohla jen "ležet na internetu".

Modely ale umí částečně i generalizovat: kombinovat naučené vzory v originální odpovědi. Důvodem může být to, že s velkým množstvím dat může být generalizovat jednodušší, než se učit vše nazepamět.

A tady už přichází další otázky: Co dělá inteligentní člověka? A co je to vlastně ta "intelligence"?

Kde se můžu naučit o jazykových modelech více?

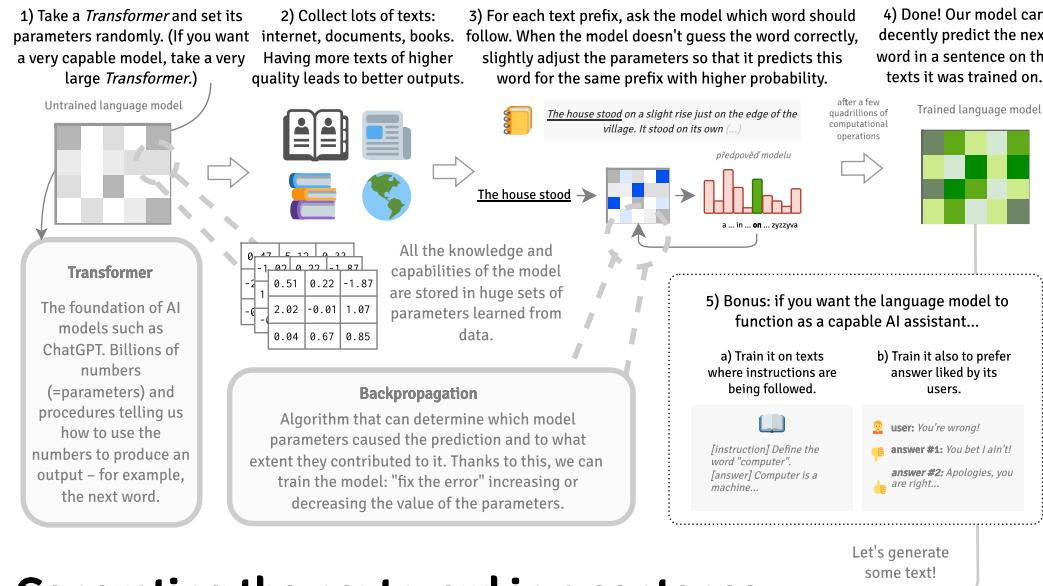


Krátká odpověď: U nás na ÚFALu!

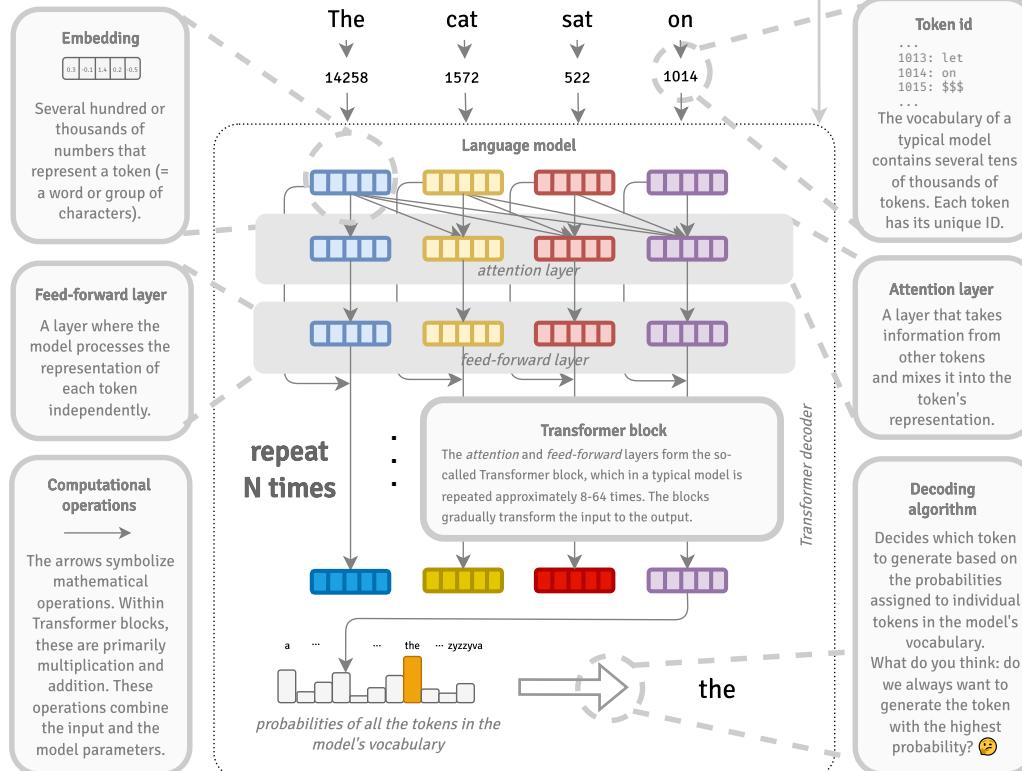
Dlouhá odpověď: Podobnými tématy se zabýváme na Ústavu Formální a Aplikované Lingvistiky (ÚFAL), což je součást Matematicko-fyzikální fakulty Univerzity Karlové.

Na naší katedře propoujdejeme znalosti o jazyce se znalostmi ze strojového učení. Učíme o tom bakalářské i magisterské předměty, jako třeba Deep Learning nebo Large Language Models.

Recipe for the model behind ChatGPT



Generating the next word in a sentence



NEW: See it all animated at <https://animatedllm.github.io>

Frequently asked questions about language models

Can language models also look up answers on the internet?

Short answer: They can't! (But someone can do it for them.)

Long answer: A language model as such cannot search the internet. It had to learn everything it can do during training.

Nowadays, however, commercial language models typically have an external search module. This module searches for information related to the query and appends it to the user input. The model then works with this information the same way as with the rest of the text. The quality of the search results can influence the generated text both positively and negatively.

Does the language model have a list of all words in all languages?

Short answer: No – but of their parts, yes!

Long answer: There are indeed too many words in all the languages combined (think of all the Czech declensions!).

Instead, we use so-called "subwords": words and their parts from which we can assemble the rest by "gluing" the parts together. We have the more frequent words in the vocabulary directly and we assemble the less common ones from multiple parts. Breaking text into subwords and then re-assembling it is the task of a specialized algorithm, the so-called tokenizer, which works independently of the language model.

Why don't models know other languages as well as English?

Short answer: Because English is the language of the internet.

Long answer: We still can't teach models language nearly as efficiently as, for example, children. Not even all of Wikipedia is enough for a model to learn the language perfectly. The model's capabilities therefore grow with the amount of texts it was trained on. And there is really a lot of English texts online!

Surprisingly, there is quite a lot texts on the internet even in Czech. But in Irish or Telugu, for example, the model will lag behind.

How energy-intensive is it to train a model? And how much energy does one query consume?

Short answer: Training is demanding, generating less so.

Long answer: It depends on the size of the model, but it's estimated that training a model with 175 billion parameters costs 1.2 GWh of energy, which corresponds to the annual consumption of 120 American households. Fortunately, models are not trained that often. Generating text from a trained model is much more efficient: an average ChatGPT response consumes approximately 0.3 Wh, which corresponds to, for example, an electric kettle turned on for 1 second.

[1] <https://blog.camilmiran.com/the-painful-singularity>
[2] <https://idact.org/how-much-energy-do-lms-consume-unveiling-the-power-behind-ai/>
[3] <https://andymasley.substack.com/p/a-cheat-sheet-for-conversations-about>

What makes large language models so intelligent?

Short answer: Such a tricky question...!

Long answer: Models can to some extent repeat what they've seen in the training data. Their smart answers might be therefore just "parroted" from someone on the internet. But to some extent, the models can also combine learned patterns in an original answer. It is because with a large amount of data, it might be simpler to learn to generalize than to memorize everything.

And here come further questions: What makes a human intelligent? And what actually is this "intelligence"?

Where can I learn more about language models?

Short answer: Here at ÚFAL!

Long answer: We deal with similar topics at the Institute of Formal and Applied Linguistics (ÚFAL), which is part of the Faculty of Mathematics and Physics at Charles University. At our department, we connect knowledge about language with knowledge from machine learning.

We teach bachelor's and master's courses about it, such as Deep Learning or Large Language Models.

