

Can we detect ChatGPT in Czech and Slovak?

Petr Šigut and Tomáš Foltýnek

ChatGPT 3.5 ▾



You

write an essay on the differences between various nlp fields



ChatGPT

Title: Exploring the Diverse Landscape of Natural Language Processing (NLP) Fields

Introduction:

Natural Language Processing (NLP) is a multidisciplinary field that encompasses various subfields, each addressing distinct challenges and applications. This essay aims to explore the differences between several NLP fields, shedding light on their unique characteristics, methodologies, and real-world implications.

1. **Machine Translation (MT):**

Machine Translation focuses on developing algorithms and models to automatically translate text or speech from one language to another. This field has evolved significantly with the rise of neural machine translation (NMT) models, which leverage deep learning techniques for improved accuracy and fluency.

AI-generated text

Threats for society

- Academic integrity
- Low-quality content
- Fabrication of reviews
- Spread of misinformation
- Reliance on a hallucinated text

Context: **When I was running through the park I met a _____**

Next word candidates:

friend

group

boy

girl

gardener

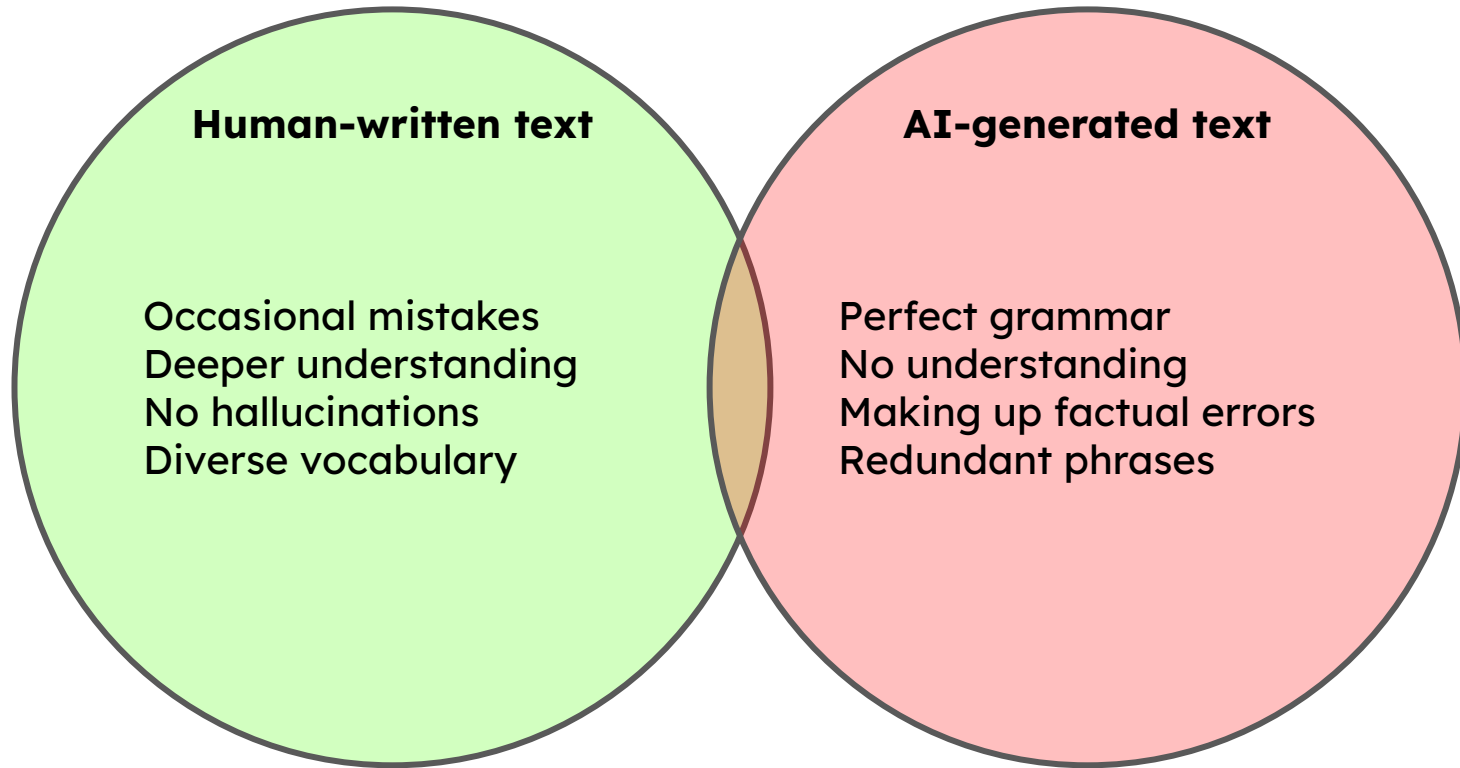
parrot

monkey

surfing

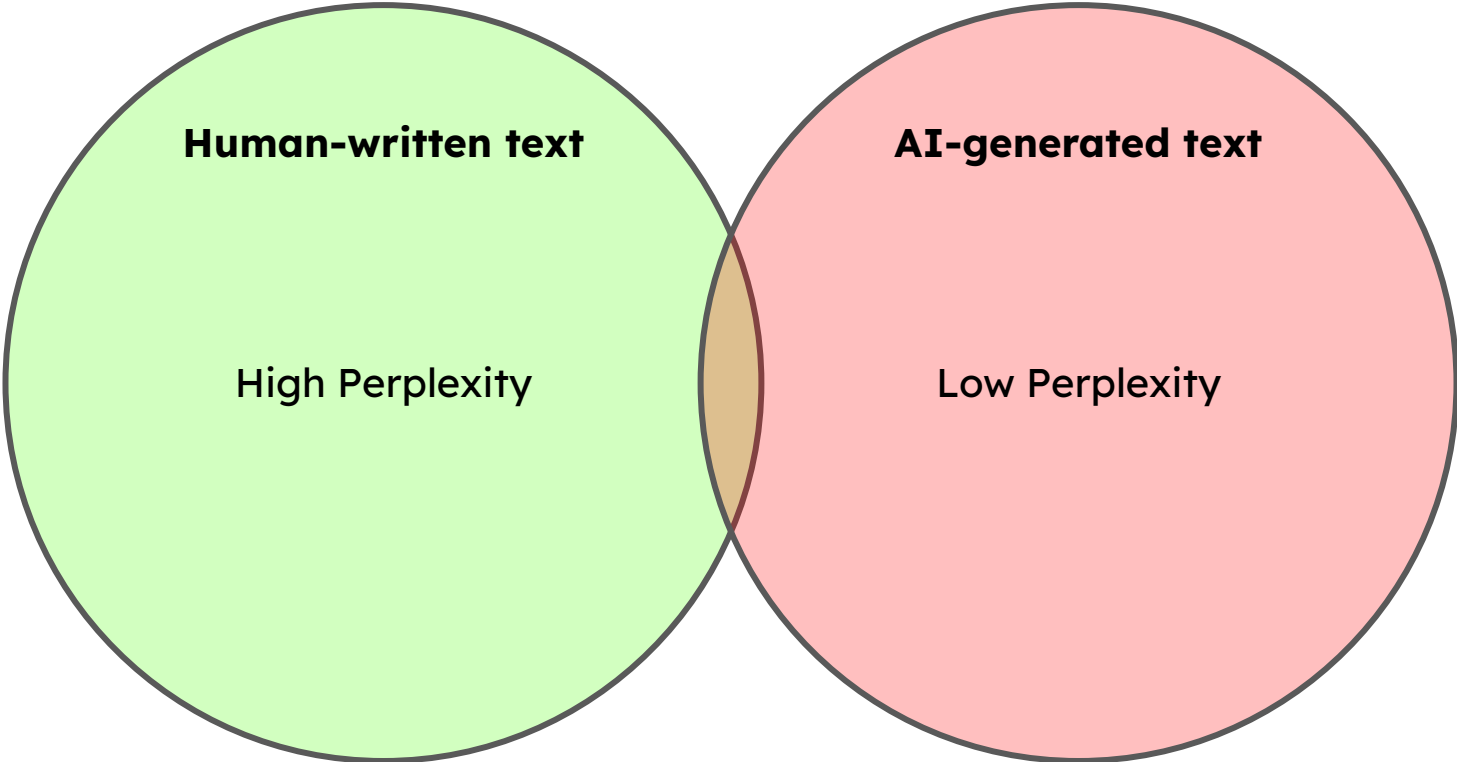
pretty

How GPTs generate text?



Differences between AI-generated and human-written text

Perplexity = the amount of randomness in text



Context: **When I was running through the park I met a _____**

- AI-detector guesses the next word given the context
- If it is often correct, the document has a low perplexity

friend

group

boy

girl

gardener

parrot

monkey

surfing

pretty

How to detect AI-generated text?

What tools did we test?



All about Compilatio AI Detector

Stay tuned about Compilatio releases

AI Detector Evaluation

Check if a text was written by an AI or not

In this trial version, your text should:

- Be written in English, French, Italian or Spanish
- Contain more than 200 characters and less than 500 characters
- Be entirely written by a human, or entirely written by an AI

When I was running in the park, I met a friend. The autumn air was crisp, and the leaves crunched beneath my feet as I followed the winding trail. As I approached a familiar bench, I saw Mark, a childhood friend I hadn't seen in ages, sitting there with a smile that mirrored my own surprise.

"Is that really you?" he exclaimed, his eyes widening in disbelief.

We exchanged a hearty handshake and settled on the bench,



420 characters - Min: 200 characters / Max: 500 characters

Start again



88% reliability

Are you satisfied with the result ?



GPT-2 Output Detector Demo

This is an online demo of the GPT-2 output detector model, based on the 🤖/Transformers implementation of RoBERTa. Enter some text in the text box; the predicted probabilities will be displayed below. The results start to get reliable after around 50 tokens.

When I was running in the park, I met a friend. The autumn air was crisp, and the leaves crunched beneath my feet as I followed the winding trail. As I approached a familiar bench, I saw Mark, a childhood friend I hadn't seen in ages, sitting there with a smile that mirrored my own surprise.

"Is that really you?" he exclaimed, his eyes widening in disbelief.

We exchanged a hearty handshake and settled on the bench,



Real

Prediction based on 92 tokens

Fake

63.68%



36.32%

Research questions

1. Can AI-detectors accurately detect Czech and Slovak content?
2. If we translate Czech and Slovak texts to English using DeepL, does the AI-detection still work?
3. Is ChatGPT3.5 easier to detect than the premium ChatGPT4?

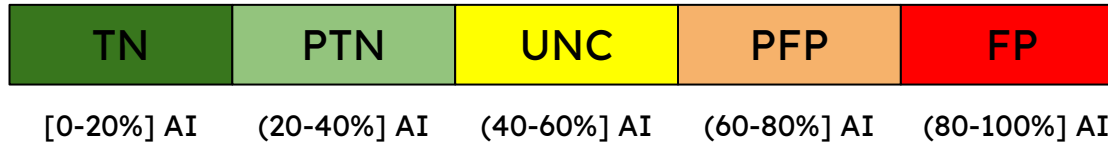
Document set

- 9 short documents in each of 15 categories
 - In total 135 documents
1. Human-written
 - a. in Czech
 - b. in Slovak
 - c. in English
 2. ChatGPT3.5 generated
 - a. in Czech
 - b. in Slovak
 - c. in English
 3. ChatGPT4 generated
 - a. in Czech
 - b. in Slovak
 - c. in English
 4. Translated to English
 - a. Human-written in Czech (1a)
 - b. Human-written in Slovak (1b)
 - c. ChatGPT3.5 generated in Czech (2a)
 - d. ChatGPT3.5 generated in Slovak (2b)
 - e. ChatGPT4 generated in Czech (3a)
 - f. ChatGPT4 generated in Slovak (3b)

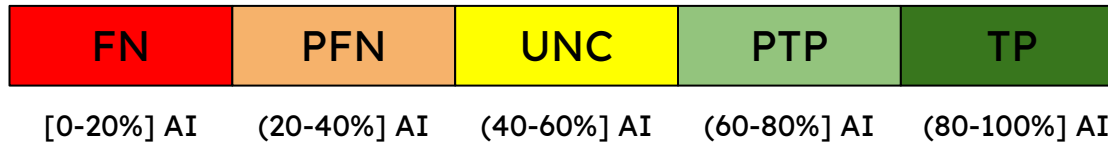
Evaluation

- Confidence score 0-100 % AI

Human-written



AI-generated



Relevant Metrics

$$\textit{Accuracy} = \frac{TP + TN + 0.5 * (PTP + PTN)}{\text{n of all documents}} * 100$$

Relevant Metrics

$$Sensitivity_{original} = \frac{TP}{TP + FN} * 100$$

Measures how often the detector correctly detects AI-generated content among the AI-generated documents

$$Sensitivity = \frac{TP + 0.5 * PTP}{n \text{ of AI-generated documents}} * 100$$

Relevant Metrics

$$Specificity_{original} = \frac{TN}{TN + FP} * 100$$

Measures how often the detector correctly detects human-written document among the human-written documents

$$Specificity = \frac{TN + 0.5 * PTN}{n \text{ of human-written documents}} * 100$$

Evaluation of the results

- 18 documents = 9 Human-written + 9 AI-generated

ChatGPT 3.5

English	Compilatio	GPT-2 Output Detector
Specificity	100 %	67 %
Sensitivity	22 %	28 %
Accuracy	61 %	47 %

Czech	Compilatio	GPT-2 Output Detector	Compilatio with translated text	GPT-2 OD with translated text
Specificity	56 %	100 %	78 %	61 %
Sensitivity	56 %	0 %	11 %	33 %
Accuracy	56 %	50 %	44 %	47 %

Slovak	Compilatio	GPT-2 Output Detector	Compilatio with translated text	GPT-2 OD with translated text
Specificity	61 %	100 %	89 %	83 %
Sensitivity	72 %	0 %	0 %	44 %
Accuracy	67 %	50 %	44 %	64 %

ChatGPT 4

English	Compilatio	GPT-2 Output Detector
Specificity	100 %	67 %
Sensitivity	0 %	0 %
Accuracy	50 %	33 %

Czech	Compilatio	GPT-2 Output Detector	Compilatio with translated text	GPT-2 OD with translated text
Specificity	56 %	100 %	78 %	61 %
Sensitivity	67 %	0 %	22 %	33 %
Accuracy	61 %	50 %	50 %	47 %

Slovak	Compilatio	GPT-2 Output Detector	Compilatio with translated text	GPT-2 OD with translated text
Specificity	61 %	100 %	89 %	83 %
Sensitivity	67 %	0 %	11 %	39 %
Accuracy	64 %	50 %	50 %	61 %

Conclusions

- Compilatio was the only multi-lingual detector
- Translating to English is effective to some extent
- ChatGPT4 is less detectable than ChatGPT3.5 in English
- In Czech and Slovak ChatGPT4 had very similar results