CTU

CZECH TECHNICAL UNIVERSITY IN PRAGUE

Mining data about people from Wikipedia using LLMs

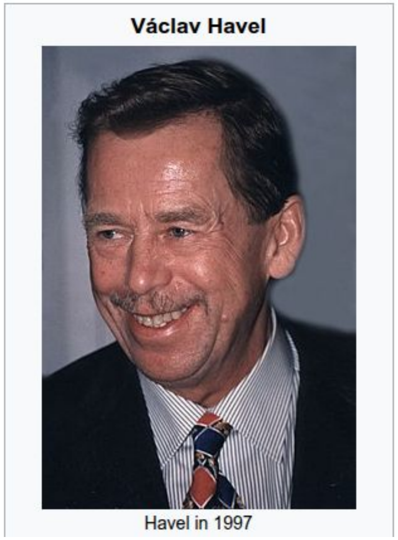
summer semester 2024/2025

Bc. Matyáš Meisner	- annotation tool
Bc. Maroš Mešter	- regex solution
Bc. Radim Průdek	- LLM-based solution
Bc. Aleš Trna	- data analysis

Abstract

The aim of this project is to utilize **Large Language Models** to automatically extract age-related information from articles about real **people on Wikipedia**. This is done in order to create a database of people's faces annotated with their age. To evaluate this method of annotation, **a ground truth annotation database** together with **a tool for manual annotation** was developed. Finally, the accuracy and coverage of the LLM-based approach was evaluated.

image of a person on Wikipedia



Václav Havel

Havel in 1997

image caption

Project goals

1. Create a “baseline“ automated solution without using AI.
2. Create a ground truth database of an adequate size (can be used as metric).
3. Create manual annotation tool for creating such database.
4. Design a LLM-based automated annotation system.
5. Compare the models and create a performance report.

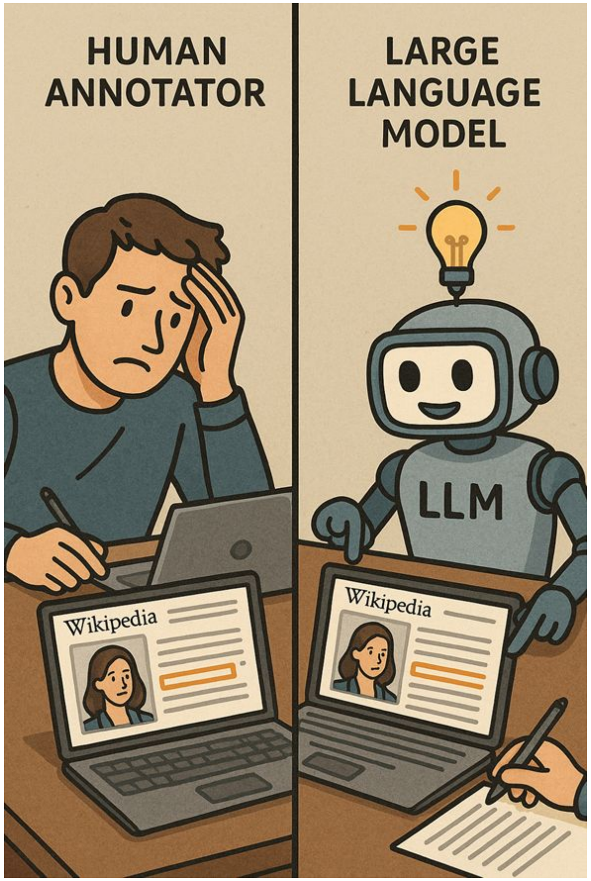
Background & Motivation

- Databases with reliable annotation are **expensive** and **time consuming.**

- Are Large language models capable of doing such tasks?

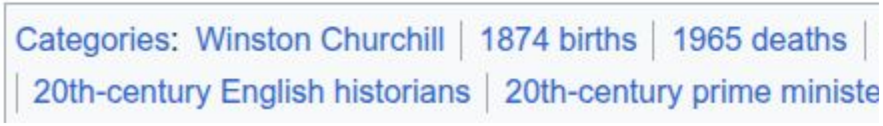
- Database of labeled images for training **an age estimating mode**l.

- Possibility of mining huge amount of free data from a large source (Wikipedia).



Baseline automated solution - Regex

- Python regex library
- extracts birth year from the *category* box on Wikipedia
- tries to extract year from image caption using:
  - r"\b[12]\d{3}\b"
  - r"\b[12]\d{3}['’]?s\b"

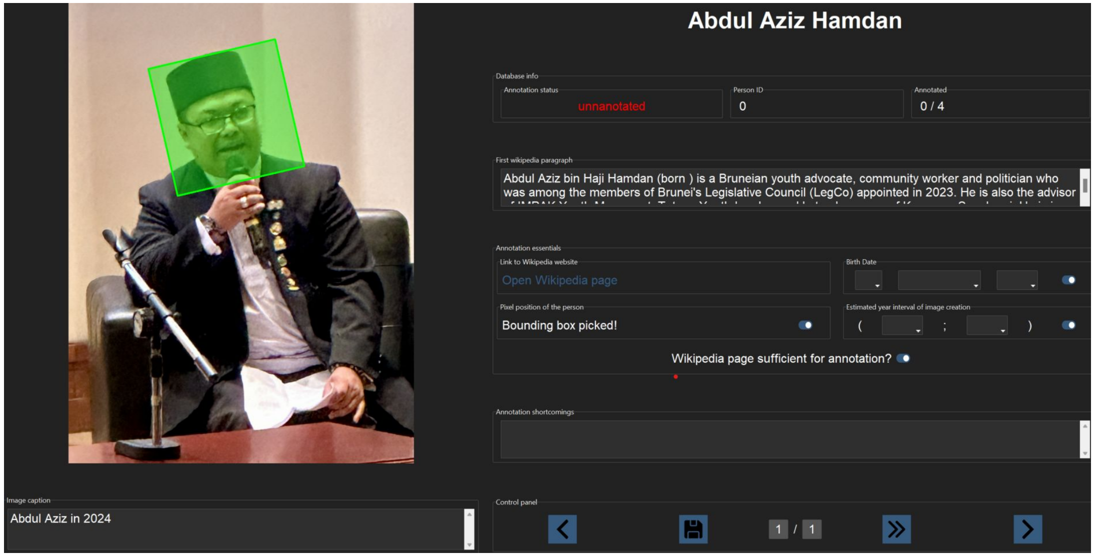


Categories: Winston Churchill | 1874 births | 1965 deaths | 20th-century English historians | 20th-century prime ministe

Churchill with Lord Halifax in 1938

Image annotation tool

- GUI for manual annotation (Python’s ttkbootstrap and tkinter)
- features:
  - selecting bounding box
  - filling annotation data
  - navigating through database
  - clickable link to Wikipedia article
  - noting annotation shortcomings



Abdul Aziz Hamdan

Database info

Annotation status: unnanotated

Person ID: 0

Annotated: 0 / 4

First wikipedia paragraph

Abdul Aziz bin Haji Hamdan (born) is a Bruneian youth advocate, community worker and politician who was among the members of Brunei's Legislative Council (LegCo) appointed in 2023. He is also the advisor

Annotation essentials

Link to Wikipedia website: Open Wikipedia page

Birth Date

Pixel position of the person: Bounding box picked!

Estimated year interval of image creation (;)

Wikipedia page sufficient for annotation?

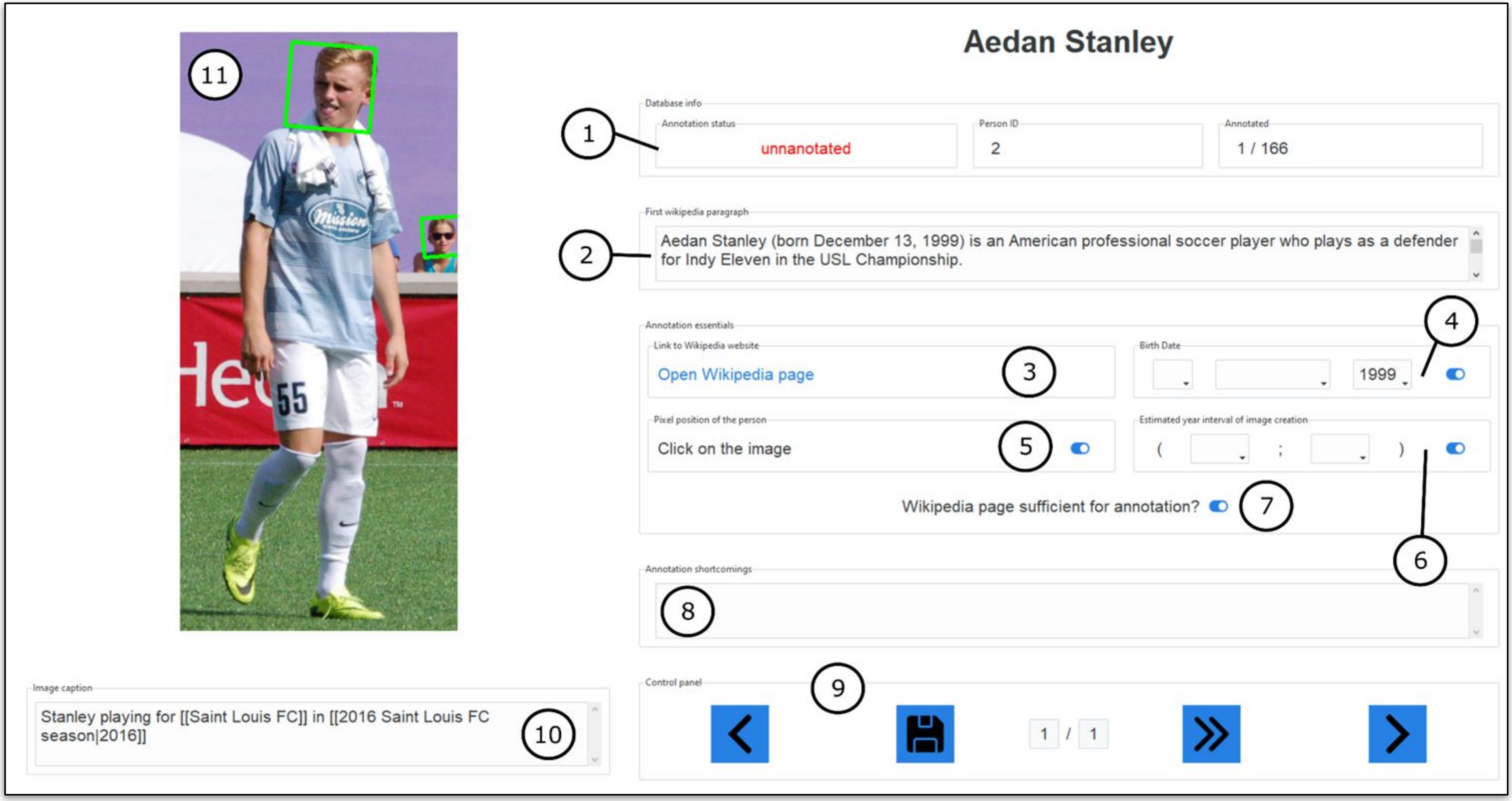
Annotation shortcomings

Image caption: Abdul Aziz in 2024

Control panel

Annotation tool (GUI) - dark mode.

Image annotation tool



Aedan Stanley

Database info

Annotation status: unnanotated

Person ID: 2

Annotated: 1 / 166

First wikipedia paragraph: Aedan Stanley (born December 13, 1999) is an American professional soccer player who plays as a defender for Indy Eleven in the USL Championship.

Annotation essentials

Link to Wikipedia website: Open Wikipedia page

Birth Date: 1999

Pixel position of the person: Click on the image

Estimated year interval of image creation: (;)

Wikipedia page sufficient for annotation?

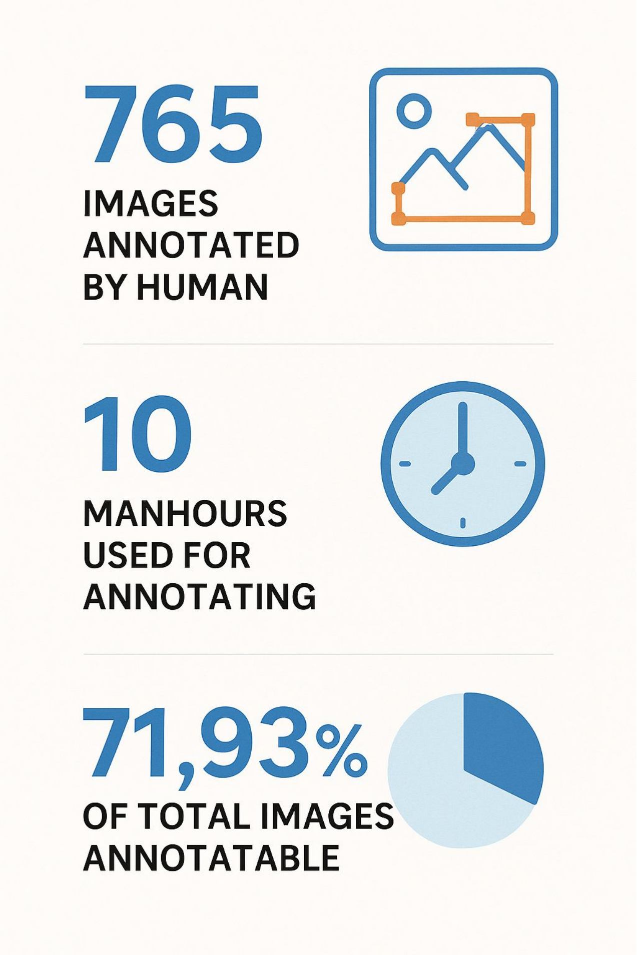
Annotation shortcomings

Control panel

Image caption: Stanley playing for [[Saint Louis FC]] in [[2016 Saint Louis FC season|2016]]

Annotation tool (GUI) - light mode.

Ground truth database



765

IMAGES ANNOTATED BY HUMAN

10

MANHOURS USED FOR ANNOTATING

71,93%

OF TOTAL IMAGES ANNOTATABLE

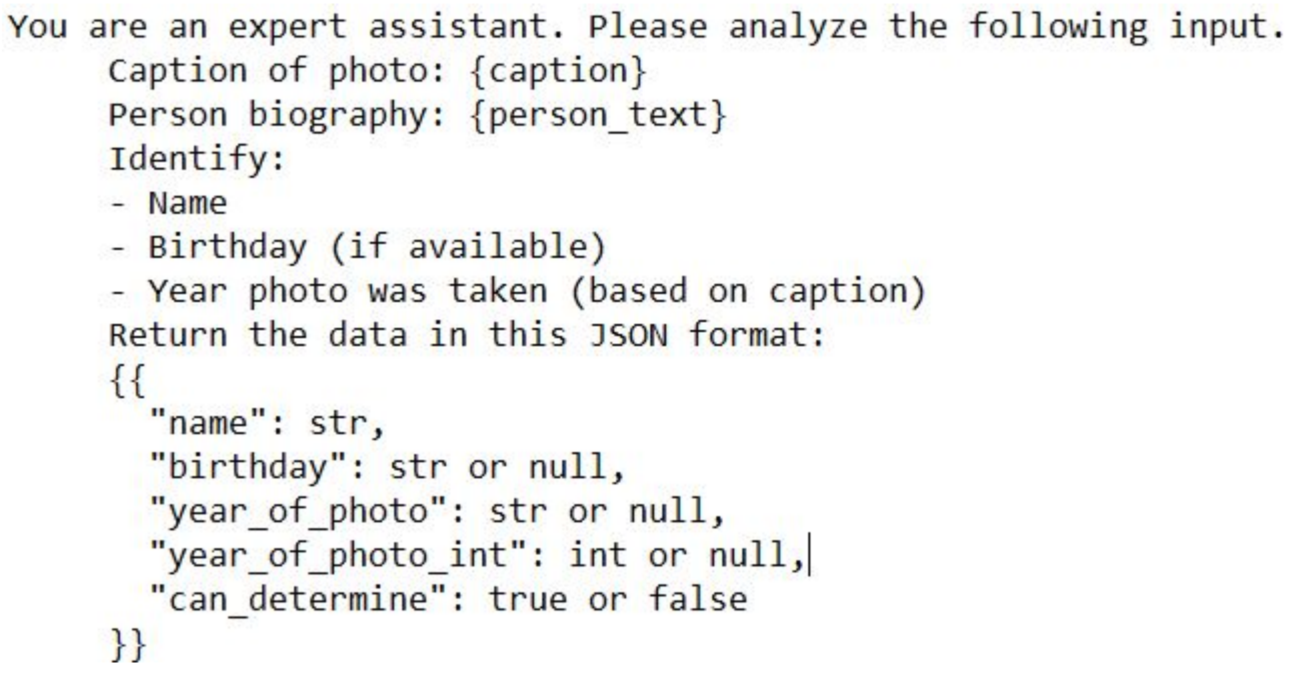
LLM – based automated solution

- Ollama - Local LLM framework.
- Returns structured outputs.
- Tested on different LLMs on RCI server

Model	Mean time per annotation [s]	Mean GPU usage [GB]
Llama3.3-70B	3.4577	43.13
DeepSeek-R1-Distill-Llama-70B	3.5558	43.08
DeepSeek-R1-Distill-Qwen-32B	2.7843	21.21
DeepSeek-R1-Distill-Qwen-14B	1.3514	10.59
DeepSeek-R1-Distill-Llama-8B	0.6689	6.29
DeepSeek-R1-Distill-Qwen-7B	0.8965	5.44

LLM – based automated solution

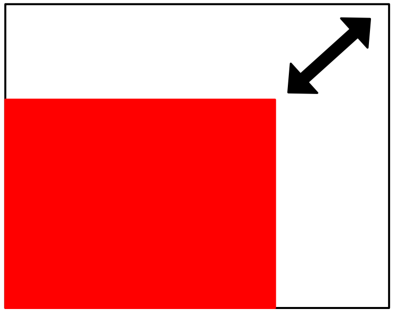
- **Six different prompts** for estimating the age of the person.
- Example prompt:



```
You are an expert assistant. Please analyze the following input.
     Caption of photo: {caption}
     Person biography: {person_text}
     Identify:
     - Name
     - Birthday (if available)
     - Year photo was taken (based on caption)
     Return the data in this JSON format:
     {{
       "name": str,
       "birthday": str or null,
       "year_of_photo": str or null,
       "year_of_photo_int": int or null,
       "can_determine": true or false
     }}
```

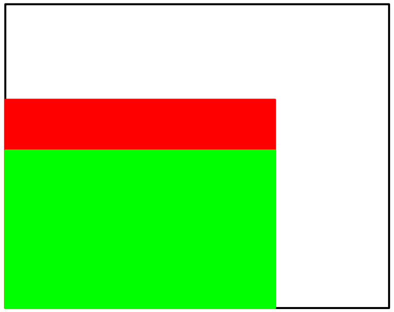
Results

Coverage of each model/prompt



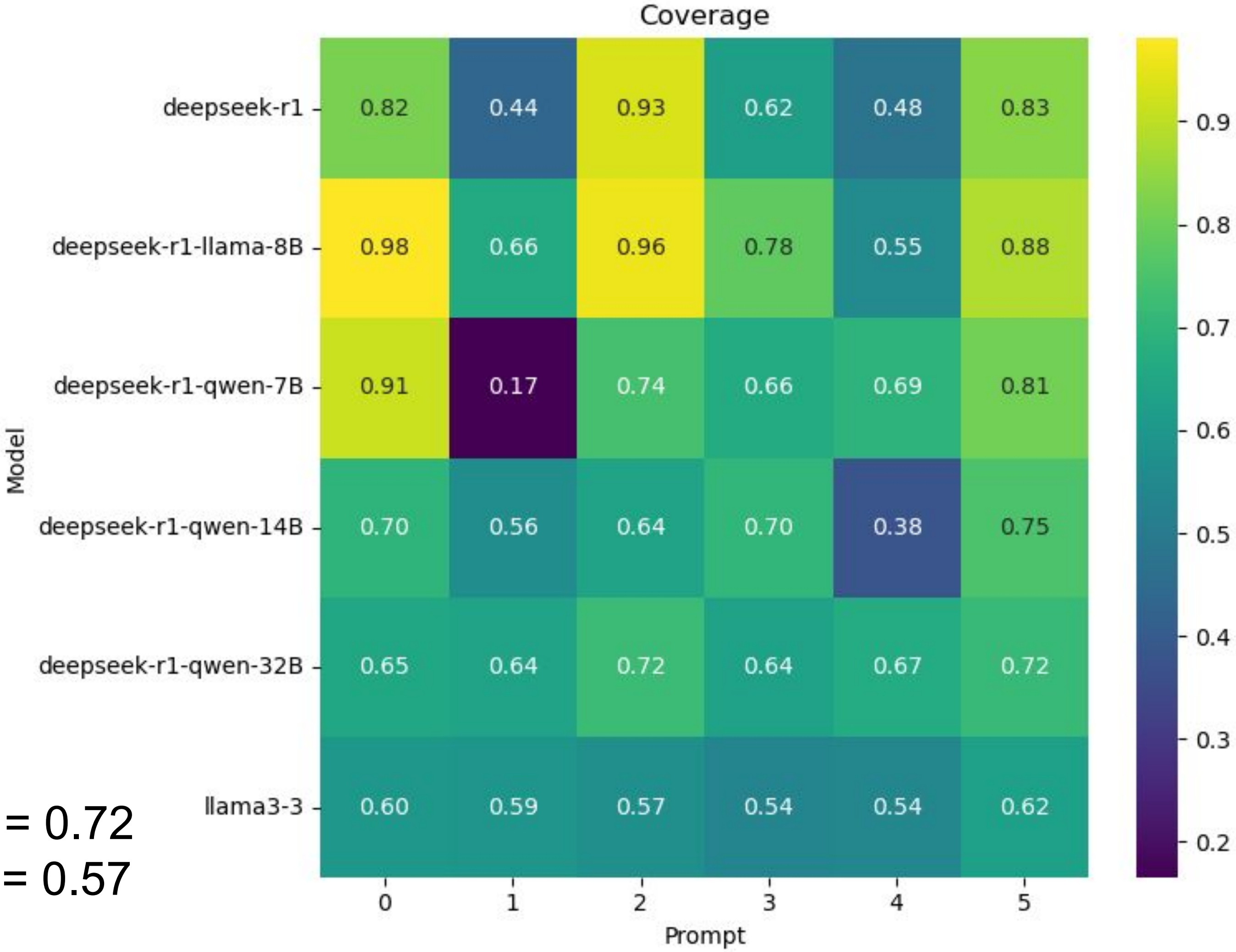
coverage

Accuracy on different quantities tested



accuracy

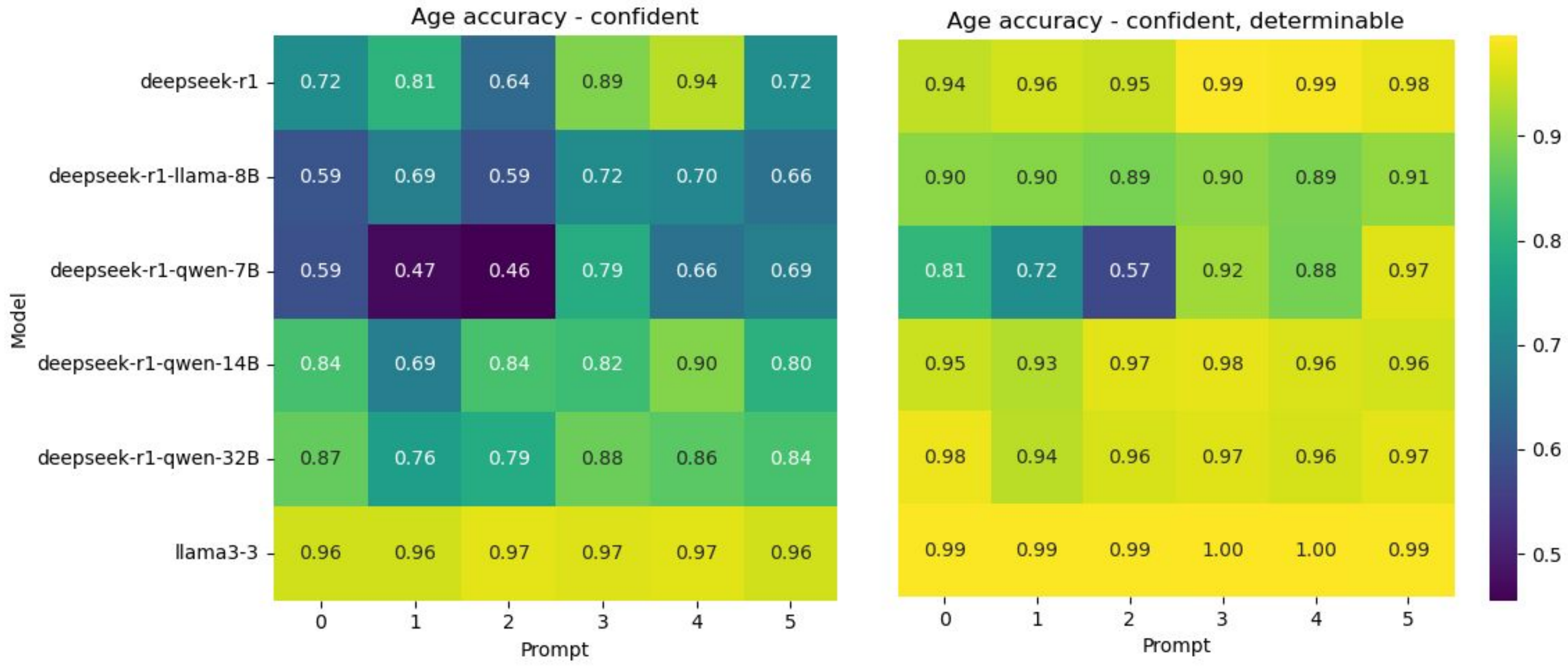
Coverage - ability to estimate age


Coverage

Model	0	1	2	3	4	5
deepseek-r1	0.82	0.44	0.93	0.62	0.48	0.83
deepseek-r1-llama-8B	0.98	0.66	0.96	0.78	0.55	0.88
deepseek-r1-qwen-7B	0.91	0.17	0.74	0.66	0.69	0.81
deepseek-r1-qwen-14B	0.70	0.56	0.64	0.70	0.38	0.75
deepseek-r1-qwen-32B	0.65	0.64	0.72	0.64	0.67	0.72
llama3-3	0.60	0.59	0.57	0.54	0.54	0.62

Human coverage = 0.72
Regex coverage = 0.57

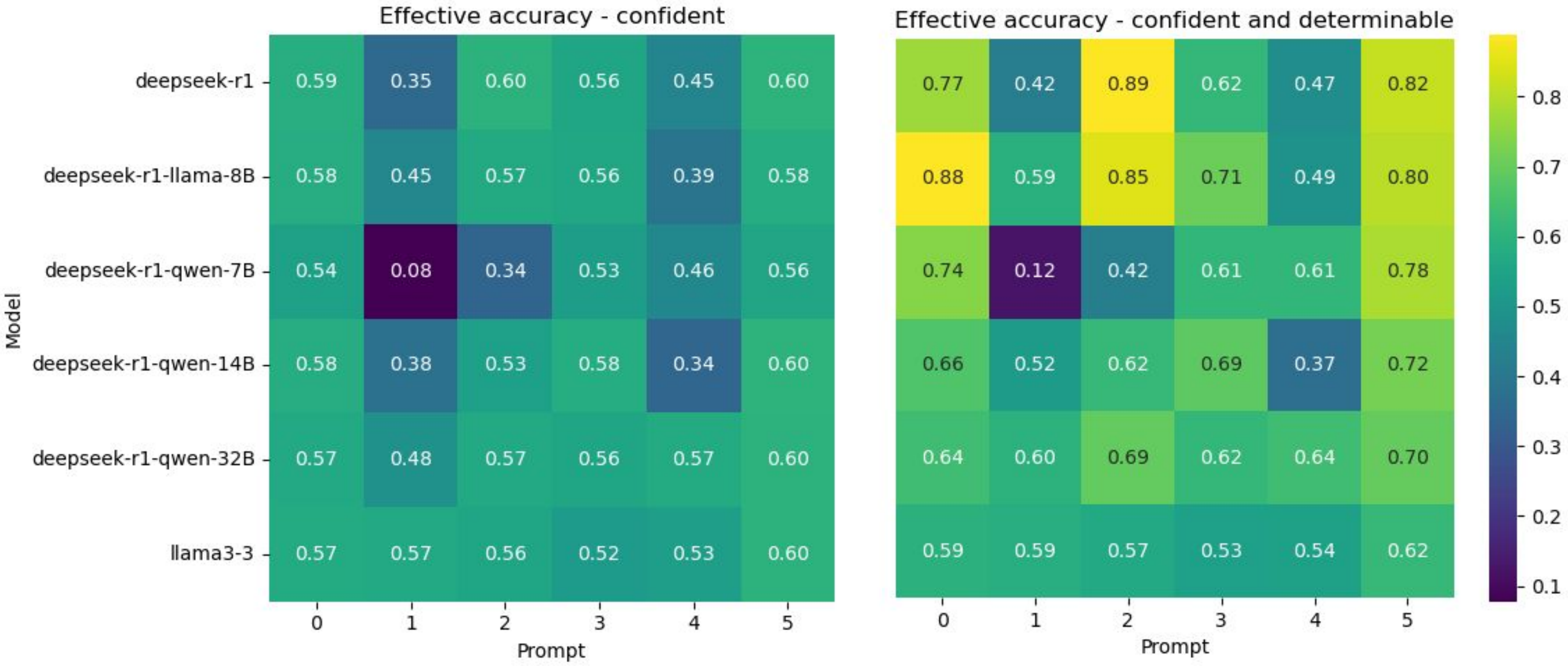
Predicted age accuracy
model is confident
age may not be determinable



Regex accuracy = 0.84

Regex accuracy = 0.98

“Effective accuracy”
product of accuracy and coverage



Regex - 0.48

Regex - 0.56

Questions?

Resources

https://rci.cvut.cz/

https://coderpad.io/wp-content/uploads/2022/04/coderpad-regex-the-complete-guide.jpg

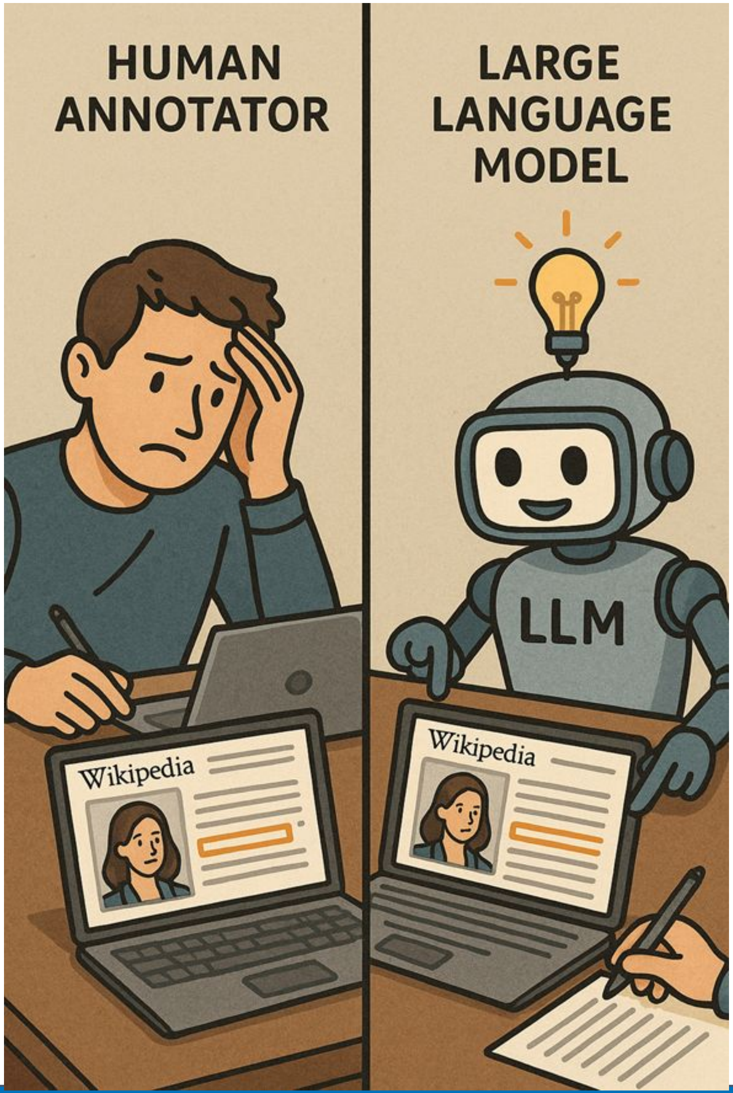
Some illustrative images were generated by OpenAI SORA model

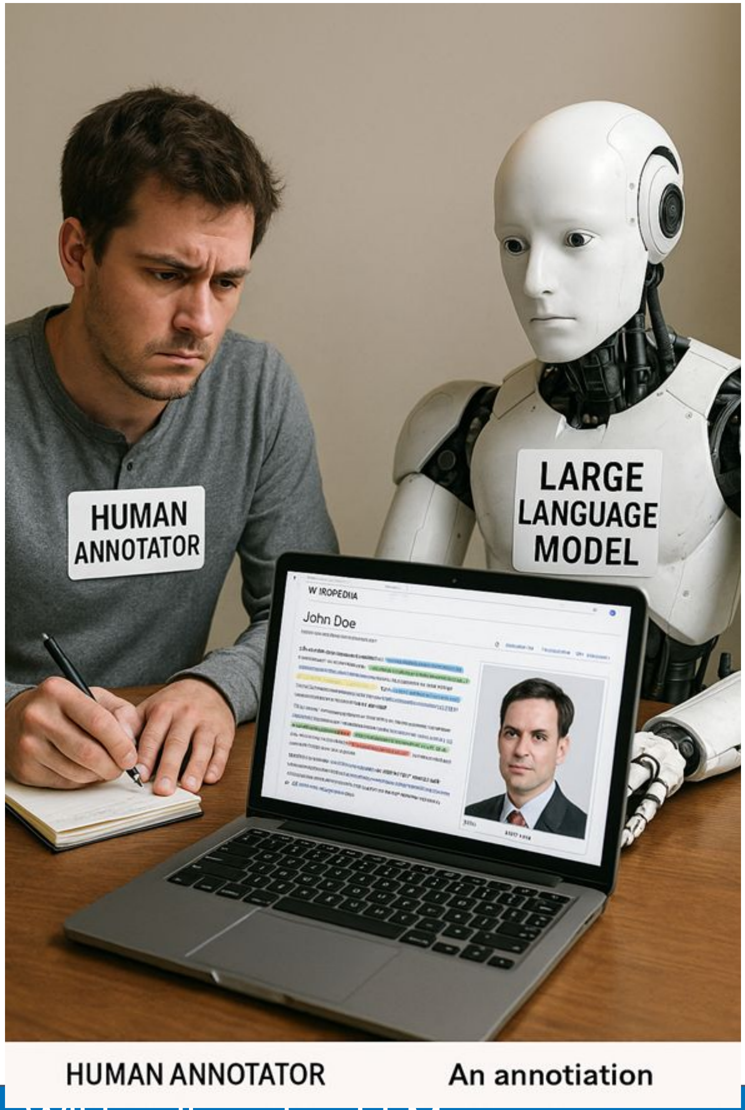
CTU

CZECH TECHNICAL UNIVERSITY IN PRAGUE

Thank you for your attention

“Create an image which shows human annotator on one side
and Large language model doing an annotation task on a person on wikipedia.
Make sure to show the advantages and struggles of both actors.“


HUMAN ANNOTATOR
LARGE LANGUAGE MODEL
LLM
Wikipedia
Wikipedia


HUMAN ANNOTATOR
LARGE LANGUAGE MODEL
W IROPEDIA
John Doe
HUMAN ANNOTATOR
An annotation