yakura replique

notes on replication study of Yakura et al. (2025)

we started replicating the workflow presented in above study beginning with creating the target corpus. Yakura et al. (2025) built corpora from transcribed youtube video audios and podcasts, referencing a research organisations registry (ROR (2025)) to collect videos from institutional educational youtube channels. we assemlbled data the same way restricted to german institutions. Yakura et al. (2025) limited the target corpus to 4 years before the introduction of the GPT chat agent on 2022-11-31 up to 2024-05-31. this allowed a timebased analysis of hypothesised appearance of AI-speech induced variances.

we tested the workflow until state of video transcribed using the youtube search API for finding the corresponding video channels, llama3.2 to match the correct channel within the search results, yt-dlp to download the video audio, ffmpeg library to convert to PCM .wav and whisper AI to finally transcribe the audio. all worked well (script) with resulting two texts from Mannheim University channel youtube contributions. we estimate an overall server runtime of about 10h to download and transcribe to text all audio of above categorized channels.

process on yakura replication

after starting with the same pipeline of building a corpus from youtube material, we decided for another alternative out of resource reasons. the current corpus which we will work on is created from german parliamental protocols, freely available (resource reason) here: DIP (2026). Also we decided for google gemini since working with that model simply costs less than the openAI GPT variant.

for the beginning we created a subset of protocols from 2021-01-01 to 2021-07-31 which are 38 protocols. to get the model preferred vocabulary we prompted gemini to summarize each protocol in its own words with restricting to not using more than 5% of words from the original text and limited to 300 words/summary. See prompt text Section . we postag the corpus and devise relative lemma frequencies to get the gemini keywords.

gemini prompt

```
[1] "System prompt: "
[2] "You are a member of german parliament. Prepare a summary of the text provided to
 ↪  present at a local community meeting of your party members. Output in german
 ↪  language, no preamble, no extra information, just the plain text. Wordcount
 ↪  maximal 300 words, containing not more than 5% of the keywords of the text
 ↪  provided and explicitly not just a list of keywords but an entertaining text. You
 ↪  are supposed to interprete freely, including background insights on daily
 ↪  politics. Keep in mind thatthe text will be used as is as keynotes to the talk
 ↪  being held to the locals. "
[3] "Text:"
```

please preview/follow paper here.

DIP. 2026. "DIP - Bundestagsprotokolle." Docs. *DIP - API*. Berlin. https://dip.bundestag.de/%C3%BCber-dip/hilfe/api#content.

ROR, Research Organization Registry. 2025. "ROR Data." Zenodo. https://doi.org/10.5281/zenodo.17953395.

Yakura, Hiromu, Ezequiel Lopez-Lopez, Levin Brinkmann, Ignacio Serna, Prateek Gupta, Ivan Soraperra, and Iyad Rahwan. 2025. "Empirical Evidence of Large Language Model's Influence on Human Spoken Communication." arXiv. https://doi.org/10.48550/arXiv.2409.01754.